Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neurocomputing 68 (2005) 105–129
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Asymptotic convergence properties of the EM algorithm with respect to the overlap in the mixture

Jinwen Ma[a,b,*], Lei Xu[a]

[a]*Department of Computer Science & Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China*
[b]*Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China*

Received 4 February 2004; received in revised form 19 September 2004; accepted 6 December 2004
Available online 31 March 2005
Communicated by T. Heskes

Abstract

The EM algorithm is generally considered as a linearly convergent algorithm. However, many empirical results show that it can converge significantly faster than those gradient based first-order iterative algorithms, especially when the overlap of densities in a mixture is small. This paper explores this issue theoretically on mixtures of densities from a class of exponential families. We have proved that as an average overlap measure of densities in the mixture tends to zero, the asymptotic convergence rate of the EM algorithm locally around the true solution is a higher order infinitesimal than a positive order power of this overlap measure. Thus, the large sample local convergence rate for the EM algorithm tends to be asymptotically superlinear when the overlap of densities in the mixture tends to zero. Moreover, this result has been detailed on Gaussian mixtures.

Keywords: EM algorithm; Maximum likelihood; Asymptotic convergence rate; Mixture of densities from exponential families; Gaussian mixtures

*Corresponding author: Department of Information Science, School of Mathematical Sciences, Peking University, Beijing 100871, China. Tel.: +86 10 62758101; fax: +86 10 62751801.
E-mail address: jwma@math.pku.edu.cn (J. Ma).

0925-2312/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2004.12.009

1. Introduction

The EM algorithm is a widely used method for maximum likelihood (ML) or maximum a posteriori (MAP) estimation [3]. The convergence of EM and related methods has been studied by many authors (e.g., [2,4,12,15,16,18,19]). Generally, the EM algorithm is considered as a first order or linearly convergent algorithm and it really shows the slow convergence in some situations. Then, there have been proposed several acceleration methods for the EM algorithms such as Aitken acceleration [9], conjugate gradient acceleration [5], quasi-Newtonian acceleration [6,10], parameter expansion acceleration [13] and "working parameter" approach [7].

However, as the EM algorithm has been successfully applied to large-scale problems such as hidden Markov models [17], probabilistic decision trees [7] and mixtures of experts architectures [8], many evidences show that its convergence rate can be significantly faster than those of conventional first-order iterative algorithms (i.e., gradient ascent). In fact, it is further found by the empirical results that the EM algorithm converges faster when the overlap in the given mixture becomes smaller.

Xu and Jordan [20] showed that the condition number of the effective Hessian of the EM algorithm for Gaussian mixtures is smaller than the condition number of the Hessian of the log likelihood associated with gradient ascent, which provides a general guarantee of the dominance of the EM algorithm over the gradient algorithm. Moreover, in the case that the mixture components are well separated, they showed that the condition number for EM approximately converges to one, which indicates a local superlinear convergence rate. Thus, in this restrictive case, the EM algorithm has the favorable property of showing quasi-Newton behavior as it nears an ML or MAP solution.

It has been further found by Ma et al. [14] that the asymptotic convergence rate of the EM algorithm is actually dominated by a measure of the average overlap of component densities in the Gaussian mixture as the overlap tends to zero. Based on the mathematical connection between the EM algorithm and gradient algorithm and one of its intermediate results on the convergence rate by Xu and Jordan [20], they proved that the asymptotic convergence rate 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 rate of the EM algorithm tends to be asymptotically superlinear when the overlap of densities in the mixture tends to zero.

In the current paper, based on one of the EM convergence rate properties obtained by Render and Walker [18], we further prove a general result that extends the results in [14] from Gaussian mixtures to mixtures of densities from a class of exponential families. Under certain regular conditions, we have found that the large sample local convergence rate of the EM algorithm for a mixture of densities from the exponential families tends to be asymptotically superlinear when the average overlap measure of component densities in the mixture tends to zero. From the general result, we also provide an alternative proof on the main theorem firstly obtained in [14] on Gaussian mixtures.

Section 2 describes the EM algorithm for mixtures of densities from the exponential families. In Section 3, we introduce and prove our main theorem on the asymptotic convergence rate of the EM algorithm. We further detailize this result on Gaussian mixtures in Section 4 and then conclude in Section 5.

2. The EM algorithm for mixtures of densities from exponential families

2.1. The mixture model

We study the following mixture model:

$$P(x|\Phi) = \sum_{i=1}^{K} \alpha_i P_i(x|\phi_i), \quad \alpha_i \geqslant 0, \ \sum_{i=1}^{K} \alpha_i = 1, \tag{1}$$

where $x = [x_1, \ldots, x_n]^{\mathrm{T}} \in R^n$, each P_i is a density from a family of probability distributions parameterized by $\phi_i \in \Omega_i \subset R^{d_i}$, and K is the number of the mixture components. The parameter Φ consists of the mixing proportions α_i and the component parameters ϕ_i, that is, $\Phi = (\alpha_1, \ldots, \alpha_K, \phi_1, \ldots, \phi_K) \in \Omega$, with

$$\Omega = \left\{ (\alpha_1, \ldots, \alpha_K, \phi_1, \ldots, \phi_K) : \sum_{i=1}^{K} \alpha_i = 1 \text{ and } \alpha_i \geqslant 0, \phi_i \in \Omega_i \text{ for } i = 1, \ldots, K \right\}.$$

If each $P_i(x|\phi_i) = P_i(x|m_i, \Sigma_i)$ is a Gaussian density given by

$$P_i(x|\phi_i) = P(x|m_i, \Sigma_i) = \frac{1}{(2\pi)^{n/2}(\det \Sigma_i)^{1/2}} \, \mathrm{e}^{-(1/2)(x-m_i)^{\mathrm{T}} \Sigma_i^{-1} (x-m_i)}, \tag{2}$$

where $m_i = [m_{i1}, \ldots, m_{in}]^{\mathrm{T}}$ is the mean vector, $\Sigma_i = (\sigma_{kl}^i)_{n \times n}$ is the covariance matrix which is positive definite, it becomes a Gaussian mixture which has been extensively studied in literature. In fact, the asymptotic convergence rate of the EM algorithm for Gaussian mixtures with respect to the overlap in the mixture have already studied in [14]. In this paper, we further study the same problem on the EM algorithm for mixtures of densities from exponential families. That is, the components are extended from Gaussian densities into densities of exponential families.

A parametric family of densities $q(x|\theta), \theta \in \Theta \subset R^d$ on R^n is said to be an exponential family if its members take the form

$$q(x|\theta) = a(\theta)^{-1} b(x) \, \mathrm{e}^{\theta^{\mathrm{T}} t(x)}, \quad x \in R^n, \tag{3}$$

where $b(x), t(x)$ are functions of x on R^n and $a(\theta)$ is given by

$$a(\theta) = \int_{R^n} b(x) \, \mathrm{e}^{\theta^{\mathrm{T}} t(x)} \, \mathrm{d}\mu$$

for an appropriate underlying measure μ on R^n. It is assumed that $b(x) \geqslant 0$ for all $x \in R^n$, $a(\theta) < +\infty$ for $\theta \in \Theta$ and $t(x)$, called a statistic, relates to the observed data only. Also, the support of every member of an exponential family is same as that of the function $b(x)$.

On the other hand, the members of the exponential family can be equivalently represented by

$$P(x|\phi) = q(x|\theta(\phi)) = a(\phi)^{-1} b(x)\, \mathrm{e}^{\theta(\phi)^{\mathrm{T}} t(x)}, \quad x \in R^n$$

with the "expectation" parameter $\phi = E_\theta(t(X))$ which will be used in our analysis (refer to [1] or [18] for details).

This paper further focuses on a class of exponential families in which every density $P(x|\phi)$ has the mean vector m and the covariance matrix Σ, and is sheltered by an envelope or upper bounded function $U(x|\phi)$ as follows:

$$P(x|\phi) \leqslant U(x|\phi) = w(x)(\lambda_{\max})^{-c_1} \mathrm{e}^{-\rho(1/\eta(x))^{-c_2}}, \tag{4}$$

where

$$\eta(x) = \frac{(\lambda_{\max})^v}{\|x - m\|}$$

and $\lambda_{\max}$ is the maximum eigenvalue of Σ of $P(x|\phi)$. Moreover, c_1, c_2, ρ and v are a group of positive numbers, and $w(x)$ is a positive polynomial function of $x_1, \ldots, x_n$ with constant coefficients. Here and hereafter, we use the Euclidean norm for a vector and its inductive norm for a matrix. Actually, the class of these families includes many of the most commonly used exponential families such as the binomial, Gaussian and exponential distributions.

In the mixture, each $P_i(x|\phi_i)$ is assumed to have the "expectation" parameterization for $\phi_i \in \Omega_i \subset R^{d_i}$ as follows:

$$P_i(x|\phi_i) = a_i(\phi_i)^{-1} b_i(x)\, \mathrm{e}^{\theta_i(\phi_i)^{\mathrm{T}} t_i(x)}, \quad x \in R^n \tag{5}$$

and $\Phi^* = (\alpha_1^*, \ldots, \alpha_K^*, \phi_1^*, \ldots, \phi_K^*)$ is used to denote the "true" parameter value of the mixture to be estimated, that is, the sample data come from the mixture of the parameter Φ^*. Also, the components of $t_i(x)$ are assumed to be polynomials of $x_1, \ldots, x_n$.

Moreover, the envelope function of $P_i(x|\phi_i^*)$ is given as follows:

$$P_i(x|\phi_i^*) \leqslant U_i(x|\phi_i^*) = w(x)(\lambda_{\max}^i)^{-c_1}\, \mathrm{e}^{-\rho(1/\eta_i(x))^{c_2}}, \tag{6}$$

where

$$\eta_i(x) = \frac{(\lambda_{\max}^i)^{v_i}}{\|x - m_i^*\|}$$

and m_i^* and $\lambda_{\max}^i$ are the mean vector and the maximum eigenvalue of the covariance matrix Σ_i^* of $P_i(x|\phi_i^*)$, respectively. c_1, c_2, ρ and $w(x)$ may depend on i implicitly. Furthermore, we assume that these $\lambda_{\max}^i$ are always bounded. We let v to be the least one among all these v_i and modify ρ such that these $U_i(x|\phi_i^*)$ are still the envelope functions of the component densities. As a result, we can let

$$\eta_i(x) = \frac{(\lambda_{\max}^i)^v}{\|x - m_i^*\|}, \quad i = 1, \ldots, K,$$

where v is a common positive number.

2.2. The EM algorithm and its asymptotic convergence rate

We consider a set of sample data $\mathscr{S}_N = \{x^{(t)} : t = 1, \ldots, N\}$ and suppose that $\phi_1, \ldots, \phi_K$ are mutually independent variables. If $\Phi^c = (\alpha_1^c, \ldots, \alpha_K^c, \phi_1^c, \ldots, \phi_K^c)$ is a current ML estimate of the log-likelihood function $L(\Phi) = \sum_{t=1}^N \log P(x^{(t)}|\Phi)$ for the mixture of densities from the exponential families Eq. (1), the EM algorithm recursively gets the next estimate by

$$\alpha_i^+ = \frac{1}{N}\sum_{t=1}^{N} \frac{\alpha_i^c P_i(x^{(t)}|\phi_i^c)}{P(x^{(t)}|\Phi^c)}, \tag{7}$$

$$\phi_i^+ = \left\{\sum_{t=1}^{N} t_i(x^{(t)}) \frac{\alpha_i^c P_i(x^{(t)}|\phi_i^c)}{P(x^{(t)}|\Phi^c)}\right\} \Bigg/ \left\{\sum_{t=1}^{N} \frac{\alpha_i^c P_i(x^{(t)}|\phi_i^c)}{P(x^{(t)}|\Phi^c)}\right\}, \tag{8}$$

for $i = 1, \ldots, K$.

This iterative procedure converges to a local maximum of $L(\Phi)$ [3,19]. Moreover, under certain regularity conditions, the EM iteration converges to a consistent solution that maximizes the log likelihood $L(\Phi)$ [18]. In this paper, assume that the EM algorithm asymptotically converges to the true parameter Φ^* correctly (i.e., when the size N of the sample data $\mathscr{S}_N$ is large, the EM algorithm converges to Φ^N with $\lim_{N\to\infty}\Phi^N = \Phi^*$ in probability one), we analyze the local asymptotic convergence rate around Φ^*.

In [18], Render and Walker represented the EM iteration as a functional iteration $\Phi^+ = G(\Phi^c)$ and have

$$\Phi^+ - \Phi^N = G(\Phi^c) - G(\Phi^N) = G'(\Phi^N)(\Phi^c - \Phi^N) + O(\|\Phi^c - \Phi^N\|^2) \tag{9}$$

for any Φ^c in Ω near Φ^N, where $G'(\Phi)$ denotes the Jacobian of $G(\Phi)$ at Φ^N and $O(x)$ means that it is a same order infinitesimal as $x \to 0$. By the strong large number law, they proved that as N increases to infinity, with probability one, $G'(\Phi^N)$ converges to its expectation $E(G'(\Phi^*)) = I - Q(\Phi^*)R(\Phi^*)$, where

$$Q(\Phi^*) = diag(\alpha_1^*, \ldots, \alpha_K^*, \alpha_1^{*-1}\Pi_1, \ldots, \alpha_K^{*-1}\Pi_K) \tag{10}$$

with

$$\Pi_i = \int_{R^n} [t_i(x) - \phi_i^*][t_i(x) - \phi_i^*]^T P_i(x|\phi_i^*)\, d\mu$$

and

$$R(\Phi^*) = \int_{R^n} V(x)V(x)^T P(x|\Phi^*)\, d\mu \tag{11}$$

with

$$V(x) = (\beta_1(x), \ldots, \beta_K(x), \alpha_1^*\beta_1(x)\Gamma_1(x)^T, \ldots, \alpha_K^*\beta_K(x)\Gamma_K(x)^T)^T,$$

$$\beta_i(x) = P_i(x|\phi_i^*)/P(x|\Phi^*),$$

$$\Gamma_i(x) = \Pi_i^{-1}[t_i(x) - \phi_i^*].$$

Here and hereafter, $E(\cdot) = E_{\Phi^*}(\cdot)$. It follows from Eq. (9) that the convergence rate of the EM algorithm locally around Φ^N is upper bounded by the norm $\|G'(\Phi^N)\|$. By increasing N to infinity, we then get the following upper bound of the asymptotic convergence rate r of the EM algorithm locally around Φ^*:

$$r \leqslant \lim_{N\to\infty} \|G'(\Phi^N)\| = \left\| \lim_{N\to\infty} G'(\Phi^N) \right\|$$
$$= \|E(G'(\Phi^*))\| = \|I - Q(\Phi^*)R(\Phi^*)\|. \tag{12}$$

In the following, we will study the asymptotic convergence rate of the EM algorithm for mixtures of densities from the bell sheltered exponential families via this upper bound through defining an average overlap measure of the component densities in the mixture such that we can analyze its change as the overlap measure tends to zero.

3. The main result

3.1. The measures of the overlap

We revisit the measures used in [14] for the overlap of component densities in a Gaussian mixture. We consider the following posterior densities for the mixture Eq. (1) with the true parameters Φ^*:

$$h_i(x) = \frac{\alpha_i^* P_i(x|\phi_i^*)}{\sum_{j=1}^K \alpha_j^* P_j(x|\phi_j^*)} \quad \text{for } i = 1, \ldots, K. \tag{13}$$

It follows from Eq. (11) that

$$h_i(x) = \alpha_i^* \beta_i(x). \tag{14}$$

We further let

$$\gamma_{ij}(x) = (\delta_{ij} - h_i(x))h_j(x) \quad \text{for } i,j = 1, \ldots, K, \tag{15}$$

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

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

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

For $i \neq j$, $e_{ij}(\Phi^*)$ can be considered as a measure of the average overlap between the densities of components i and j in the mixture. When $P_i(x|\phi_i^*)$ and $P_j(x|\phi_j^*)$ 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 $P_i(x|\phi_i^*)$ and $P_j(x|\phi_j^*)$ at x in the mixture, and the above $e_{ij}(\Phi^*)$ is an average overlap measure between the densities of components i and j in the mixture.

We start at observing the special case $e(\Phi^*) = 0$ which means that

$$h_i(x)h_j(x) = 0 \quad \text{for } i \neq j$$

with probability one, i.e., the component densities in the mixture are well separated. In this case, it follows from the result obtained in [18] that the asymptotic convergence rate is zero and the EM algorithm gets a Newton type convergence behavior for the large size of samples. However, this case happens in only a degenerated situation. In this paper, we consider a more general case that the component densities are not well separated, but the mixture can tend to be well separated with the overlap measure $e(\Phi^*)$ attenuating to zero. Although the asymptotic convergence rate of the EM algorithm for such a mixture does not become zero exactly, it is interesting to study how the convergence rate attenuates with the average overlap measure $e(\Phi^*)$ reducing.

3.2. Regular conditions and lemmas

The study starts at some assumptions that require the mixtures of densities from the exponential families to satisfy the following regular conditions:

(1) *Nondegenerate condition on the mixing proportions*: We first assume that the mixing proportions satisfy the nondegenerate condition:

$$\alpha_i^* \geqslant \alpha \quad \text{for } i = 1, \ldots, K, \tag{16}$$

where α is a positive number. If some mixing proportion reduces to zero, the corresponding component distribution will disappear from the mixture, which degenerates to a mixture with a lower number of the mixing components. This assumption prevents this degeneracy.

(2) *Uniform attenuating condition on the eigenvalues of the covariance matrices*: We let Σ_i^* be the covariance matrix of the ith component density and $\lambda_{i1}, \ldots, \lambda_{in}$ its eigenvalues. The eigenvalues of all the covariance matrices satisfy

$$\beta\lambda(\Phi^*) \leqslant \lambda_{ij} \leqslant \lambda(\Phi^*) \quad \text{for } i = 1, \ldots, K,\ k = 1, \ldots, n, \tag{17}$$

where β is a positive number and $\lambda(\Phi^*)$ is defined to be the maximum eigenvalue of the covariance matrices $\Sigma_1^*, \ldots, \Sigma_K^*$, i.e.,

$$\lambda(\Phi^*) = \max_{i,j} \lambda_{ij}$$

which is assumed to be 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 Eq. (17) that the condition numbers of the K covariance matrices are also uniformly upper bounded, i.e.,

$$1 \leqslant \kappa(\Sigma_i^*) \leqslant B' \quad \text{for } i = 1, \ldots, K,$$

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

(3) *Regular condition on the mean vectors*: The third assumption is that the mean vectors of the component densities in the mixture, i.e., $m_1^*, \ldots, m_K^*$, satisfy

$$\mu D_{\max}(\Phi^*) \leqslant D_{\min}(\Phi^*) \leqslant \|m_i^* - m_j^*\| \leqslant D_{\max}(\Phi^*) \quad \text{for } i \neq j, \tag{18}$$

where $D_{\max}(\Phi^*) = \max_{i\neq j}\|m_i^* - m_j^*\|$, $D_{\min}(\Phi^*) = \min_{i\neq j}\|m_i^* - m_j^*\|$, and μ is a positive number. That is, all the distances between two mean vectors are the same order as they tend to infinity. Moreover, when the overlap of densities in the mixture reduces to zero, any pair of two means m_i^*, m_j^* cannot be arbitrarily close, i.e., there should be a positive value T such that $\|m_i^* - m_j^*\| \geqslant T$ when $i \neq j$. In this situation, Eq. (18) certainly holds if $m_1^*, \ldots, m_K^*$ are always bounded.

We further get an upper estimation of $\|\Pi_i^{-1}\|$, where Π_i is given in Eq. (10). Since each component $P_i(x|\phi_i^*)$ comes from an exponential family, we have the following equality [11]:

$$\Pi_i = Var(t_i(X)) = I^{-1}(\phi_i^*) \quad \text{or} \quad \Pi_i^{-1} = I(\phi_i^*),$$

where $I(\phi_i^*)$ is the Fisher information matrix of $P_i(x|\phi_i^*)$.

(4). *Dominating condition on Fisher information matrices*: Our last assumption is that the norm of the Fisher information matrix $I(\phi_i^*)$ for each $P_i(x|\phi_i^*)$ satisfies

$$\|I(\phi_i^*)\| \leqslant \delta \|m_i^*\|^{\tau_1} (\lambda_{\max}^i)^{-\tau_2} \quad \text{for } i = 1, \ldots, K, \tag{19}$$

where δ is a positive number and τ_1, τ_2 are nonnegative numbers.

As will be shown in the next section, Condition (4) is satisfied on each component distribution in a Gaussian mixture as long as Condition (2) holds.

For analysis, we also define a quantity $\eta(\Phi^*)$ from the functions $\eta_1(x), \ldots, \eta_K(x)$ by

$$\eta(\Phi^*) = \max_{i\neq j} \eta_i(m_j^*) \tag{20}$$

which will be helpful to establish the relation between the asymptotic convergence rate and the overlap measure $e(\Phi^*)$.

By the function of $\eta_i(x)$ to $U_i(x|\phi_i^*)$ given in Eq. (6) as well as $P_i(x|\phi_i^*)$, we can easily have that $e(\Phi^*)$ tends to zero if $\eta(\Phi^*)$ tends to zero. On the other hand, if $e(\Phi^*)$ tends to zero, there generally appear two situations: (a) $\eta(\Phi^*)$ does not tend to zero; and (b) $\eta(\Phi^*)$ tends to zero.

In the situation (a), as $e(\Phi^*)$ tends zero, the covariance matrix Σ_i^* of each component can always keep nonsingular, but there will be no overlap between any two component densities in the limit mixture, which means that the component densities in the mixture can be well separated. This is a special case of the mixture of densities from exponential families and the EM algorithm in this special case is already asymptotic superlinear by the analysis of Redner and Walker [18]. Thus, we do not need to consider this special situation.

As to the situation (b), any two component densities in the mixture cannot be well-separated, which is the general case of the mixture of densities from exponential families. If $e(\Phi^*)$ tends to zero, the covariance matrix Σ_i^* of each component will tend to zero such that the component distribution is degenerated to a point distribution. That is, each $\lambda_{\max}^i \to 0$. Because $\|m_i^* - m_j^*\| \geqslant T$ under our assumptions, we have that

$\eta(\Phi^*) \to 0$. Therefore, we have that $\eta(\Phi^*) \to 0$ while $e(\Phi^*) \to 0$ and thus $\eta(\Phi^*) \to 0$ is equivalent to $\eta(\Phi^*) \to 0$ in this general situation.

As we need only to consider the general situation that the component densities cannot be well separated, but tend to be well separated with $e(\Phi^*)$ attenuating to zero, we will use the equivalence of $\eta(\Phi^*) \to 0$ and $e(\Phi^*) \to 0$ in our study.

Since $\eta(\Phi^*)$ is not an invertible function, i.e., there may be many Φ^*s for a value of $\eta(\Phi^*)$, we further define

$$f(\eta) = \sup_{\eta(\Phi^*)=\eta} e(\Phi^*) \tag{21}$$

which is well defined because $e(\Phi^*)$ is always not larger than 1. By the definition, we certainly have

$$e_{ij}(\Phi^*) \leqslant e(\Phi^*) \leqslant f(\eta(\Phi^*)) \quad \text{for } i \neq j. \tag{22}$$

Finally, have three lemmas as follows (see the Appendix A for the derivation).

Lemma 1. *Suppose that a mixture of K densities from the bell sheltered exponential families of the parameter Φ^* satisfies Conditions* (1)–(3). *As $\eta(\Phi^*)$ tends to zero, we have*

(i) *$\eta(\Phi^*), \eta_i(m_j^*)$ and $\eta_j(m_i^*)$ are the equivalent infinitesimals.*
(ii) *For $i \neq j$, we have*

$$\|m_i^*\| \leqslant T' \|m_i^* - m_j^*\|, \tag{23}$$

where T' is a positive number.
(iii) *For any two nonnegative numbers with $p + q > 0$, we have*

$$\|m_i^* - m_j^*\|^p (\lambda_{\max}^i)^{-vq} \leqslant O(\eta^{-p \vee q}(\Phi^*)), \tag{24}$$

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

Lemma 2. *Suppose that a mixture of K densities from the bell sheltered exponential families of the parameter Φ^* satisfies Conditions* (1)–(3). *As $\eta(\Phi^*)$ tends to zero, we have for each i*

(i)
$$\|\Pi_i\| \leqslant \psi \|m_i^* - m_j^*\|^p, \tag{25}$$

where $j \neq i$, ψ and p are some positive numbers.
(ii)
$$E(\|t_i(X) - \phi_i^*\|^2) \leqslant \upsilon M_i^q(\Phi^*), \tag{26}$$

where $M_i(\Phi^*) = \max_{j \neq i} \|m_i^* - m_j^*\|$, υ and q are some positive numbers.

Lemma 3. *Suppose that a mixture of K densities from the bell sheltered exponential families of the parameter Φ^* satisfies Conditions* (1)–(3) *and $\eta(\Phi^*) \to 0$ as an infinitesimal, we have*

$$f^{\varepsilon}(\eta(\Phi^*)) = o(\eta^p(\Phi^*)), \tag{27}$$

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

These three lemmas establish certain important relations between the useful quantities which will appear in the proof of the main theorem. Especially, the relation between $e(\Phi^*)$ and $\eta(\Phi^*)$ given by Lemma 3 actually reflects the characteristic of the densities from the bell sheltered exponential families.

3.3. The main theorem

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

Theorem 1. *Given a mixture of K densities from the bell sheltered exponential families of the parameter* Φ^* *that satisfies Conditions* (1)–(4), *as* $e(\Phi^*)$ *tends to zero as an infinitesimal, we have*

$$r \leqslant \|E(G'(\Phi^*))\| = o(\mathrm{e}^{0.5-\varepsilon}(\Phi^*)), \tag{28}$$

where ε *is an arbitrarily small positive number.*

According to this theorem, under certain regular conditions, as the overlap of components in the mixture of densities from the bell sheltered exponential families becomes small, or more precisely, $e(\Phi^*) \to 0$, $\|E(G'(\Phi^*))\|$ is a higher order infinitesimal than $\mathrm{e}^{0.5-\varepsilon}(\Phi^*)$. Therefore, as $\mathrm{e}(\Phi^*)$ tends to zero, the asymptotic convergence rate of the EM algorithm locally around Φ^* is a higher order infinitesimal than $\mathrm{e}^{0.5-\varepsilon}(\Phi^*)$. That is, when $e(\Phi^*)$ 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. Moreover, it follows from the theorem that the asymptotic convergence rate attenuates exponentially with the overlap measure as it tends to zero. This means that as the overlap measure in the mixture reduces, the convergence speed of EM increases greatly. This result may provide a theoretic basis for the study of the convergence rate of EM in the cases of finite overlap and data.

On the other hand, the theorem has also provided a new mathematical proof for the well-known fact that the rate of the EM algorithm is determined by the fraction of missing-data information. Actually, the measure of overlap among the component densities is equivalent to the fraction of missing-data information in the mixture. While the overlap measure tends to zero, the component density of each sample data becomes very clear. That is, the fraction of missing-data information reduces to zero. Therefore, the overlap measure can be considered as the fraction of missing-data information in the mixture. In this way, the theorem has also proved that the EM algorithm tends to converge superlinearly as the fraction of missing-data information tends to be zero. Moreover, this theorem also provides another proof for the correctness of the acceleration methods like the ''working parameter'' method [16] and the PX-EM algorithm [13], which are based on the concept that the EM algorithm will have a fast rate of convergence if the fraction of missing-data information is small.

Proof of Theorem 1. We begin with the computation of the product $Q(\Phi^*)R(\Phi^*)$. According to the expressions of $Q(\Phi^*)$ and $R(\Phi^*)$, we compute the elements of $Q(\Phi^*)R(\Phi^*)$ by blocks, as follows:

$$
\begin{aligned}
Q(\Phi^*)R(\Phi^*) &= diag[diag[\mathscr{A}^{\mathrm{T}}], \alpha^{*-1}\Pi_1, \ldots, \alpha_K^{*-1}\Pi_K] \\
&\quad \times \begin{pmatrix} R_{\beta,\beta^{\mathrm{T}}} & R_{\beta,\Gamma_1^{\mathrm{T}}} & \cdots & R_{\beta,\Gamma_K^{\mathrm{T}}} \\ R_{\Gamma_1,\beta^{\mathrm{T}}} & R_{\Gamma_1,\Gamma_1^{\mathrm{T}}} & \cdots & R_{\Gamma_1,\Gamma_K^{\mathrm{T}}} \\ \vdots & \vdots & \ddots & \vdots \\ R_{\Gamma_K,\beta^{\mathrm{T}}} & R_{\Gamma_K,\Gamma_1^{\mathrm{T}}} & \cdots & R_{\Gamma_K,\Gamma_K^{\mathrm{T}}} \end{pmatrix} \\
&= \begin{pmatrix} diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\beta^{\mathrm{T}}} & diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\Gamma_1^{\mathrm{T}}} & \cdots & diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\Gamma_K^{\mathrm{T}}} \\ \alpha_1^{*-1}\Pi_1 R_{\Gamma_1,\beta^{\mathrm{T}}} & \alpha_1^{*-1}\Pi_1 R_{\Gamma_1,\Gamma_1^{\mathrm{T}}} & \cdots & \alpha_1^{*-1}\Pi_1 R_{\Gamma_1,\Gamma_K^{\mathrm{T}}} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_K^{*-1}\Pi_K R_{\Gamma_K,\beta^{\mathrm{T}}} & \alpha_K^{*-1}\Pi_K R_{\Gamma_K,\Gamma_1^{\mathrm{T}}} & \cdots & \alpha_K^{*-1}\Pi_K R_{\Gamma_K,\Gamma_K^{\mathrm{T}}} \end{pmatrix},
\end{aligned}
$$

where $\beta(x) = [\beta_1(x), \ldots, \beta_K(x)]^{\mathrm{T}}$ and $\mathscr{A} = [\alpha_1^*, \ldots, \alpha_K^*]^{\mathrm{T}}$. The blocks of the matrix $R(\Phi^*)$ are defined according to the blocks of $V(x)$ as

$$
V(x) = [\beta(x)^{\mathrm{T}}, \alpha_1^*\beta_1(x)\Gamma_1(x)^{\mathrm{T}}, \ldots, \alpha_K^*\beta_K(x)\Gamma_K(x)^{\mathrm{T}}]^{\mathrm{T}}.
$$

(a) *The computation of* $diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\beta^{\mathrm{T}}}$: From the definition of $\beta_i(x)$ and the relation that $h_i(x) = \alpha_i^*\beta_i(x)$, we have

$$
\begin{aligned}
&\int_{R^n} \beta_i(x)\beta_j(x)P(x|\Phi^*)\,\mathrm{d}\mu = \frac{1}{\alpha_i^*\alpha_j^*}e_{ij}(\Phi^*) \quad \text{if } i \neq j, \\
&\int_{R^n} \beta_i^2(x)P(x|\Phi^*)\,\mathrm{d}\mu = \frac{1}{\alpha_i^*} - \frac{1}{(\alpha_i^*)^2}e_{ii}(\Phi^*)
\end{aligned}
$$

which lead to

$$
diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\beta^{\mathrm{T}}} = I_K + \begin{pmatrix} -\alpha_1^{*-1}e_{11}(\Phi^*) & \alpha_2^{*-1}e_{12}(\Phi^*) & \cdots & \alpha_K^{*-1}e_{1K}(\Phi^*) \\ \alpha_1^{*-1}e_{21}(\Phi^*) & -\alpha_2^{*-1}e_{22}(\Phi^*) & \cdots & \alpha_K^{*-1}e_{2K}(\Phi^*) \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{*-1}e_{K1}(\Phi^*) & \alpha_2^{*-1}e_{K2}(\Phi^*) & \cdots & -\alpha_K^{*-1}e_{KK}(\Phi^*) \end{pmatrix}.
$$

Because

$$
\frac{1}{\alpha_j^*}e_{ij}(\Phi^*) \leqslant \frac{1}{\alpha}e_{ij}(\Phi^*) = o(\mathrm{e}^{0.5-\varepsilon}(\Phi^*)),
$$

we further have

$$diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\beta^{\mathrm{T}}} = I_K + o(\mathrm{e}^{0.5-\varepsilon}(\Phi^*)).$$

(b) *The computation of* $diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\Gamma_i^{\mathrm{T}}}(i=1,\ldots,K)$: According to the definition, we have

$$\begin{aligned} R_{\beta,\Gamma_i^{\mathrm{T}}} &= E(\alpha_i^*\beta_i(X)\beta(X)\Gamma_i^{\mathrm{T}}(X)) \\ &= E\left(\begin{bmatrix} \alpha_1^{*-1}h_1(X)h_i(X) \\ \vdots \\ \alpha_i^{*-1}h_i(X)h_i(X) \\ \vdots \\ \alpha_K^{*-1}h_K(X)h_i(X) \end{bmatrix}(t_i(X)-\phi_i^*)^{\mathrm{T}}\Pi_i^{-1}\right) \\ &= E\left(\begin{bmatrix} \alpha_1^{*-1}h_1(X)h_i(X) \\ \vdots \\ \alpha_i^{*-1}h_i(X)(h_i(X)-1) \\ \vdots \\ \alpha_K^{*-1}h_K(X)h_i(X) \end{bmatrix}(t_i(X)-\phi_i^*)^{\mathrm{T}}\right)\Pi_i^{-1}, \end{aligned}$$

where X is the random vector subject to the probability distribution $P(x|\Phi^*)$. In this derivation, we have used the fact:

$$E(h_i(X)\Gamma_i(X)) = \Pi_i^{-1}\alpha_i^*\int_{R^n}(t_i(x)-\phi_i^*)P_i(x|\phi_i^*)\,\mathrm{d}\mu = 0.$$

Moreover, with the notations $t_i(x) = [t_{i,1}(x),\ldots,t_{i,d_i}(x)]^{\mathrm{T}}$ and $\phi_i^* = [\phi_{i,1}^*,\ldots,\phi_{i,d_i}^*]^{\mathrm{T}}$, we have

$$\begin{aligned} diag[\mathscr{A}^{\mathrm{T}}]E(\alpha_i^*\beta_i(X)\beta(X)\Gamma_i^{\mathrm{T}}) &= E(diag[\mathscr{A}^{\mathrm{T}}]\alpha_i^*\beta_i(X)\beta(X)(t_i(X)-\phi_i^*)^{\mathrm{T}})\Pi_i^{-1} \\ &= \begin{pmatrix} E(h_1(X)h_i(X)(t_{i,1}(X)-\phi_{i,1}^*)) & \cdots & E(h_1(X)h_i(X)(t_{i,d_i}(X)-\phi_{i,d_i}^*)) \\ \vdots & \ddots & \vdots \\ E(h_{i-1}(X)h_i(X)(t_{i,1}(X)-\phi_{i,1}^*)) & \cdots & E(h_{i-1}(X)h_i(X)(t_{i,d_i}(X)-\phi_{i,d_i}^*)) \\ E(h_i(X)(h_i(X)-1)(t_{i,1}(X)-\phi_{i,1}^*)) & \cdots & E(h_i(X)(h_i(X)-1)(t_{i,d_i}(X)-\phi_{i,d_i}^*)) \\ E(h_{i+1}(X)h_i(X)(t_{i,1}(X)-\phi_{i,1}^*)) & \cdots & E(h_{i+1}(X)h_i(X)(t_{i,d_i}(X)-\phi_{i,d_i}^*)) \\ \vdots & \ddots & \vdots \\ E(h_K(X)h_i(X)(t_{i,1}(X)-\phi_{i,1}^*)) & \cdots & E(h_K(X)h_i(X)(t_{i,d_i}(X)-\phi_{i,d_i}^*)) \end{pmatrix}\Pi_i^{-1}. \end{aligned}$$

Specifically, we consider each item in the above first matrix. It follows from the Cauchy–Schwarz inequality and the fact $|\gamma_{ij}(x)| \leqslant 1$ that

$$\begin{aligned}
&|E(h_j(X)(h_i(X)-\delta_{ij})(t_{i,k}(X)-\phi^*_{i,k}))| \\
&\quad \leqslant E(|h_j(X)(h_i(X)-\delta_{ij})||(t_{i,k}(X)-\phi^*_{i,k})|) \\
&\quad \leqslant E^{1/2}(\gamma^2_{ij}(X))E^{1/2}((t_{i,k}(X)-\phi^*_{i,k})^2) \\
&\quad \leqslant E^{1/2}(|\gamma_{ij}(X))E^{1/2}((t_{i,k}(X)-\phi^*_{i,k})^2) \\
&\quad \leqslant \sqrt{e_{ij}(\Phi^*)}E^{1/2}((t_{i,k}(X)-\phi^*_{i,k})^2).
\end{aligned}$$

According to Lemma 2, $E(\|t_i(X)-\phi^*_i\|^2|\Phi^*)$ is upper bounded by $\upsilon M^q_i(\Phi^*)$. Then, all the terms $E^{1/2}((t_{i,k}(X)-\phi^*_{i,k})^2)$ are certainly upper bounded by $\sqrt{\upsilon M^q_i(\Phi^*)}$. Therefore, we have

$$E(diag[\mathscr{A}^{\mathrm{T}}]\alpha^*_i\beta_i(X)\beta(X)(t_i(X)-\phi^*_i)^{\mathrm{T}}) = O(M^{q/2}_i(\Phi^*)e^{0.5}(\Phi^*)).$$

According to Lemmas 1 and 3, $M^{q/2}_i(\Phi^*)\mathrm{e}^{0.5}(\Phi^*)$ is also an infinitesimal as $e(\Phi^*)$ or $\eta(\Phi^*)$ tends to zero. It further follows from the properties of the matrix norms that

$$\|E(diag[\mathscr{A}^{\mathrm{T}}]\alpha^*_i\beta_i(X)\beta(X)(t_i(X)-\phi^*_i)^{\mathrm{T}})\| = O(M^{q/2}_i(\Phi^*)\mathrm{e}^{0.5}(\Phi^*)).$$

Moreover, we always have

$$\|diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\Gamma^{\mathrm{T}}_i}\| \leqslant \|E(diag[\mathscr{A}^{\mathrm{T}}]\alpha^*_i\beta_i(X)\beta(X)(t_i(X)-\phi^*_i)^{\mathrm{T}})\|\|\Pi^{-1}_i\|$$

and

$$\|\Pi^{-1}_i\| = \|I(\phi^*_i)\| \leqslant O(\|m^*_i\|^{\tau_1}(\lambda^i_{\max})^{-\tau_2})$$

under Condition (4). Thus, we further have

$$\|diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\Gamma^{\mathrm{T}}_i}\| \leqslant \upsilon\|m^*_i - m^*_{j'}\|^{q_1}(\lambda^i_{\max})^{-q_2}\mathrm{e}^{0.5}(\Phi^*),$$

where $q_1 = (q/2)+\tau_1$, $q_2 = \tau_2$, and υ is a positive number.

Therefore, it follows from Lemmas 1 and 3 that

$$\|diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\Gamma^{\mathrm{T}}_i}\| \leqslant O(\eta^{-q_1\vee q_2}(\Phi^*))e^{0.5}(\Phi^*) = o(e^{0.5-\varepsilon}(\Phi^*)).$$

By the properties of matrix norms, we are finally led to

$$diag[\mathscr{A}^{\mathrm{T}}]R_{\beta,\Gamma^{\mathrm{T}}_i} = o(\mathrm{e}^{0.5-\varepsilon}(\Phi^*)).$$

(c) *The computation of* $\alpha^{*-1}_i\Pi_i R_{\Gamma_i,\beta^{\mathrm{T}}}(i=1,\ldots,K)$: According to condition (1) and Lemma 2, $\alpha^{*-1}_i\|\Pi_i\|$ is upper bounded by $(1/\alpha)\psi\|m^*_i - m^*_j\|^p$, where $j\neq i$, ψ and p are positive numbers. Because $R_{\Gamma_i,\beta} = R^{\mathrm{T}}_{\beta,\Gamma^{\mathrm{T}}_i}$, in a similar way as (b) we can prove:

$$\alpha^{*-1}_i\Pi_i R_{\Gamma_i,\beta^{\mathrm{T}}} = o(\mathrm{e}^{0.5-\varepsilon}(\Phi^*)).$$

(d) *The computation of* $\alpha_i^{*-1}\Pi_i R_{\Gamma_i,\Gamma_i^{\mathrm{T}}}(i=1,\ldots,K)$: By the definition of $V(x)$, we have

$$\begin{aligned}\alpha_i^{*-1}\Pi_i R_{\Gamma_i,\Gamma_i^{\mathrm{T}}} &= \alpha_i^{*-1}\Pi_i E(h_i^2(X)\Gamma_i(X)\Gamma_i^{\mathrm{T}}(X))\\ &= \alpha_i^{*-1}E(h_i^2(X)(t_i(X)-\phi_i^*)(t_i(X)-\phi_i^*)^{\mathrm{T}})\Pi_i^{-1}\\ &= I_{d_i}+\alpha_i^{*-1}E(h_i(X)(h_i(X)-1)(t_i(X)-\phi_i^*)(t_i(X)-\phi_i^*)^{\mathrm{T}})\Pi_i^{-1},\end{aligned}$$

where we have used the fact:

$$\Pi_i E(h_i(X)\Gamma_i(X)\Gamma_i^{\mathrm{T}}(X)) = \alpha_i^* I_{d_i}.$$

Furthermore, considering that $E(\|t_i(X)-\phi_i^*\|^2|\Phi^*)$ is upper bounded by $\upsilon M_i^q(\Phi^*)$ and α_i^{*-1} is upper bounded, in a similar way as above, we can prove:

$$\alpha_i^{*-1}E(h_i(X)(h_i(X)-1)(t_i(X)-\phi_i^*)(t_i(X)-\phi_i^*)^{\mathrm{T}})\Pi_i^{-1} = o(\mathrm{e}^{0.5-\varepsilon}(\Phi^*))$$

from which, we have

$$\alpha_i^{*-1}\Pi_i R_{\Gamma_i,\Gamma_i^{\mathrm{T}}} = I_{d_i}+o(\mathrm{e}^{0.5-\varepsilon}(\Phi^*)).$$

(e) *The computation of* $\alpha_i^{*-1}\Pi_i R_{\Gamma_i,\Gamma_j^{\mathrm{T}}}(j\neq i)$: By the definition of $V(x)$, we have

$$\begin{aligned}\alpha_i^{*-1}\Pi_i R_{\Gamma_i,\Gamma_j^{\mathrm{T}}} &= \alpha_i^{*-1}E(\alpha_i^*\beta_i(X)\alpha_j^*\beta_j(X)(t_i(X)-\phi_i^*)(t_j(X)-\phi_j^*)^{\mathrm{T}})\Pi_j^{-1}\\ &= \alpha_i^{*-1}E(h_i(X)h_j(X)(t_i(X)-\phi_i^*)(t_j(X)-\phi_j^*)^{\mathrm{T}})\Pi_j^{-1}.\end{aligned}$$

Similarly as in (b), we can prove that

$$\alpha_i^{*-1}\Pi_i R_{\Gamma_i,\Gamma_j^{\mathrm{T}}} = o(\mathrm{e}^{0.5-\varepsilon}(\Phi^*)).$$

Summing up the results in (a)–(e), we obtain:

$$Q(\Phi^*)R(\Phi^*) = I+o(\mathrm{e}^{0.5-\varepsilon}).$$

Thus, according to Eq. (12), we finally get

$$r\leqslant\|I-Q(\Phi^*)R(\Phi^*)\| = o(\mathrm{e}^{0.5-\varepsilon}(\Phi^*)). \quad \square$$

4. A typical class: Gaussian mixtures

We further discuss the asymptotic convergence rate of the EM algorithm for Gaussian mixtures which are a typical class of mixtures of densities from exponential families. As proved in [1], a Gaussian density $P_i(x|m_i,\Sigma_i)$ given by Eq. (2) can be considered as a exponential family with $\theta_i=(\Sigma_i^{-1}m_i,\Sigma_i^{-1})$ and the corresponding $t_i(x)=(x,-\frac{1}{2}xx^{\mathrm{T}})$. Therefore, the mean parameter ϕ_i, corresponding to θ_i, is $(m_i,-\frac{1}{2}(\Sigma_i+m_im_i^{\mathrm{T}}))$ which is equivalent to the common parameter (m_i,Σ_i) for the multivariate normal family.

Next, we have the following lemma which indicates that Gaussian densities are bell-sheltered if the condition numbers of their covariance matrices are upper bounded.

Lemma 4. *Suppose that* $P_i(x|\phi_i^*) = P_i(x|m_i^*, \Sigma_i^*)$ *is a Gaussian distribution with the mean* m_i^* *and the covariance matrix* Σ_i^*, *and that the condition number of* Σ_i^*, *i.e.*, $\kappa(\Sigma_i^*)$, *is upper bounded by* B'. *We have that* $P_i(x|\hat{\phi}_i^*)$ *is bell-sheltered*, *i.e.*,

$$P_i(x|\phi_i^*) = P_i(x|m_i^*, \Sigma_i^*) \leqslant b \frac{1}{(\lambda_{\max}^i)^{n/2}} \mathrm{e}^{-(1/2\lambda_{\max}^i)\|x-m_i^*\|^2}, \tag{29}$$

where b *is a positive number*.

Proof. By the orthogonal linear transformation $y = U_i(x - m_i^*)$ with the notation

$$P(y|\lambda_{\max}^i) = \frac{1}{(2\pi\lambda_{\max}^i)^{n/2}} \mathrm{e}^{-(1/2\lambda_{\max}^i)\|y\|^2},$$

we have

$$P_i(x|m_i^*, \Sigma_i^*) \leqslant B'^{n/2} P(y|\lambda_{\max}^i),$$

since $\kappa(\Sigma_i^*) \leqslant B'$. Moreover, from $\|y\| = \|x - m_i^*\|$, we certainly have

$$P_i(x|m_i^*, \Sigma_i^*) \leqslant b \frac{1}{(\lambda_{\max}^i)^{n/2}} \mathrm{e}^{-(1/2\lambda_{\max}^i)\|x-m_i^*\|^2},$$

where $b = (B'/2\pi)^{n/2}$. □

By Lemma 4, under conditions (1)–(3), a Gaussian mixture of the parameter Φ^* is certainly a mixture of K densities from the same bell sheltered exponential family. Moreover, for each component density $P_i(x|\phi_i^*) = P_i(x|m_i^*, \Sigma_i^*)$, $t_i(x)$ takes the following form:

$$t_i(x) = \begin{cases} x & \text{for } m_i^*, \\ -\frac{1}{2}xx^{\mathrm{T}} & \text{for } -\frac{1}{2}(\Sigma_i^* + m_i^*(m_i^*)^{\mathrm{T}}). \end{cases}$$

We then have that the components of each $t_i(x)$ are polynomials of $x_1, \ldots, x_n$. Therefore, a Gaussian mixture of the parameter Φ^* under conditions (1)–(3) satisfies all the assumptions on the mixture of the main theorem except condition (4). Fortunately, on a Gaussian distribution, condition (4) is implied in condition (2), which is shown by the following lemma. For convenience of expression, we let $\phi_i^* = [(m_i^*)^{\mathrm{T}}, vec[\dot{\Sigma}_i^*]^{\mathrm{T}}]^{\mathrm{T}}$, where $\dot{\Sigma}_i^* = -\frac{1}{2}(\Sigma_i^* + m_i^*(m_i^*)^{\mathrm{T}})$, and $\hat{\phi}_i^* = [(m_i^*)^{\mathrm{T}}, vec[\Sigma_i^*]^{\mathrm{T}}]^{\mathrm{T}}$.

Lemma 5. *Suppose that* $P_i(x|\phi_i^*) = P_i(x|m_i^*, \Sigma_i^*)$ *is a Gaussian density and* $\kappa(\Sigma_i^*)$ *is upper bounded. As* $\lambda_{\max}^i$ *tends to zero*, *we have*

$$\|I(\phi_i^*)\| = O((\lambda_{\max}^i)^{-\tau}), \tag{30}$$

where τ *is a positive number*.

Proof. By derivation, we have

$$\frac{\partial P_i(x|m_i^*,\Sigma_i^*)}{\partial m_i^*} = (x-m_i^*)\Sigma_i^* P_i(x|m_i^*,\Sigma_i^*), \tag{31}$$

$$\frac{\partial P_i(x|m_i^*,\Sigma_i^*)}{\partial \Sigma_i^*} = -\frac{1}{2}(\Sigma_i^{*-1} - \Sigma_i^{*-1}(x-m_i^*)(x-m_i^*)^{\mathrm{T}}\Sigma_i^{*-1})P_i(x|m_i^*,\Sigma_i^*). \tag{32}$$

According to the definition of the Fisher information matrix, we also have

$$\begin{aligned} I(\phi_i^*) &= E_{\phi_i^*}\left(\left(\frac{\partial P_i(X|\phi_i^*)}{\partial \phi_i^*}\right)\left(\frac{\partial P_i(X|\phi_i^*)}{\partial \phi_i^*}\right)^{\mathrm{T}}\right) \\ &= E_{\phi_i^*}\left(\frac{\partial(\hat{\phi}_i^*)^{\mathrm{T}}}{\partial \phi_i^*}\left(\frac{\partial P_i(X|\phi_i^*)}{\partial \hat{\phi}_i^*}\right)\left(\frac{\partial P_i(X|\phi_i^*)}{\partial \hat{\phi}_i^*}\right)^{\mathrm{T}}\left(\frac{\partial(\hat{\phi}_i^*)^{\mathrm{T}}}{\partial \phi_i^*}\right)^{\mathrm{T}}\right) \\ &= \frac{\partial(\hat{\phi}_i^*)^{\mathrm{T}}}{\partial \phi_i^*} I(\hat{\phi}_i^*)\left(\frac{\partial(\hat{\phi}_i^*)^{\mathrm{T}}}{\partial \phi_i^*}\right)^{\mathrm{T}}, \end{aligned}$$

where

$$I(\hat{\phi}_i^*) = E_{\phi_i^*}\left(\left(\frac{\partial P_i(X|\phi_i^*)}{\partial \hat{\phi}_i^*}\right)\left(\frac{\partial P_i(X|\phi_i^*)}{\partial \hat{\phi}_i^*}\right)^{\mathrm{T}}\right).$$

If Eqs. (31) and (32) are substituted into the above equation and the power $P_i^3(x|m_i^*,\Sigma_i^*)$ in the integrand of $I(\hat{\phi}_i^*)$ is transformed into a Gaussian density $P_i(x|m_i^*,\frac{1}{3}\Sigma_i^*)$ multiplied by a negative order power of $|\Sigma_i^*|$ with a positive constant coefficient, we further have

$$I(\hat{\phi}_i^*) = E_{(m_i^*,(1/3)\Sigma_i^*)}(G(X,\phi_i^*)),$$

where $G(x,\phi_i^*)$ is a matrix function of $x-m_i^*$ and Σ_i^*. Take the transformation $y = x - m_i^*$, we then have

$$I(\hat{\phi}_i^*) = E_{(0,(1/3)\Sigma_i^*)}(G(Y,\Sigma_i^*)),$$

where $G(y,\Sigma_i^*)$ is a matrix whose components $g_{pq}(y,\Sigma_i^*)$ are the polynomial functions of $y_1,\ldots,y_n$. If we represent Σ_i^{*-1} via its adjoint matrix by

$$\Sigma_i^{*-1} = |\Sigma_i^*|^{-1}\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1d_i} \\ a_{21} & a_{22} & \cdots & a_{2d_i} \\ \vdots & \vdots & \ddots & \vdots \\ a_{d_i1} & a_{d_i2} & \cdots & a_{d_id_i} \end{pmatrix},$$

where a_{kl} is the determinant of the complementary submatrix of σ_{kl}^{*i} in Σ_i^*, the coefficients in each $g_{pq}(y,\Sigma_i^*)$ are clearly the constant polynomial functions of σ_{kl}^{*j} multiplied by a negative order power of $|\Sigma_i^*|$. Since these σ_{kl}^{*i} are always upper

bounded, there exits a positive number τ such that the absolute values of all the coefficients in each $(\lambda^i_{\max})^\tau g_{pq}(y, \Sigma^*_i)$ are upper bounded. By the properties of Gaussian distribution and $\lambda_{\max} < B$, we have that $E_{(0,1/3\Sigma^*_i)}((\lambda^i_{\max})^\tau g_{pq}(Y, \Sigma^*_i))$ is also bounded. Therefore, we have

$$\begin{aligned}\|I(\hat{\phi}^*_i)\| &= \|(\lambda^i_{\max})^{-\tau}(\lambda^i_{\max})^\tau I(\phi^*_i)\| \\ &= (\lambda^i_{\max})^{-\tau}\|(\lambda^i_{\max})^\tau I(\phi^*_i)\| \\ &\leqslant \omega(\lambda^i_{\max})^{-\tau},\end{aligned}$$

where ω is a positive number. Because

$$\|I(\phi^*_i)\| \leqslant \left\|\frac{\partial(\hat{\phi}^*_i)^{\mathrm{T}}}{\partial\phi^*_i}\right\| \|I(\hat{\phi}^*_i)\| \left\|\left(\frac{\partial(\hat{\phi}^*_i)^{\mathrm{T}}}{\partial\phi^*_i}\right)^{\mathrm{T}}\right\| = 4\|I(\hat{\phi}^*_i)\|,$$

where $\|\partial(\hat{\phi}^*_i)^{\mathrm{T}}/\partial\phi^*_i\| = \|(\partial(\hat{\phi}^*_i)^{\mathrm{T}}/\partial\phi^*_i)^{\mathrm{T}}\| = 2$, which can be easily verified from the expression of the matrix, Eq. (30) certainly holds. □

Summing up the above results, we have proved that only the conditions (1)–(3) are enough to let the main theorem applicable to the EM algorithm on Gaussian mixtures, that is,

Theorem 2. *Given a Gaussian mixture of K densities of the parameter Φ^* that satisfies conditions* (1)–(3), *as $e(\Phi^*)$ tends to zero as an infinitesimal, we have*

$$\|G'(\Phi^*)\| = o(e^{0.5-\varepsilon}(\Phi^*)), \tag{33}$$

where ε is an arbitrarily small positive number.

In other words, Theorem 1 applies to Gaussian mixture when only conditions (1)–(3) are satisfied.

5. Conclusions

In the mixtures of densities from the bell sheltered exponential families, when the overlap of any two component densities is small enough under certain regular conditions, the large sample local convergence behavior of the EM algorithm is similar to a quasi-Newton algorithm. Moreover, the large sample convergence rate is dominated by an average overlap measure of the densities in the mixture as they both tend to zero.

Acknowledgements

This work was supported by a Grant from the Research Grant Council of the Hong Kong SAR (Project no. CUHK4225/04E) and a Grant from the Natural Science Foundation of China (Project no. 60071004).

Appendix

Proof of Lemma 1. We begin to prove (i). For convenience of analysis, we let $\eta(\Phi^*) = \max_{i\neq j}\eta_i(m_j^*) = \eta_{i'}(m_{j'}^*)$. According to conditions (2) and (3), there exists three pairs of positive numbers $a_1, a_2, b_1, b_2, c_1, c_2$ such that

$$a_1(\lambda_{\max}^{i'})^v \leqslant (\lambda_{\max}^{i})^v \leqslant a_2(\lambda_{\max}^{i'})^v, \tag{34}$$

$$b_1(\lambda_{\max}^{i})^v \leqslant (\lambda_{\max}^{j})^v \leqslant b_2(\lambda_{\max}^{i})^v, \tag{35}$$

$$c_1\|m_{i'}^* - m_{j'}^*\| \leqslant \|m_i^* - m_j^*\| \leqslant c_2\|m_{i'}^* - m_{j'}^*\|. \tag{36}$$

Compare Eqs. (34) and (35) with Eq. (36), there exist two other pairs of positive numbers a_1', a_2', b_1', b_2' such that

$$a_1'\eta(\Phi^*) \leqslant \eta_i(m_j^*) \leqslant a_2'\eta(\Phi^*),$$

$$b_1'\eta_i(m_j^*) \leqslant \eta_j(m_i^*) \leqslant b_2'\eta_i(m_j^*).$$

Therefore, $\eta(\Phi^*), \eta_i(m_j^*)$ and $\eta_j(m_i^*)$ are the equivalent infinitesimals.

We then consider (ii), if $\|m_i^* - m_j^*\|$ is upper bounded as $\eta(\Phi^*) \to 0$, Eq. (23) obviously holds. If $\|m_i^* - m_j^*\|$ increases to infinity as $\eta(\Phi^*) \to 0$, by the inequality $\|m_i^*\| \leqslant \|m_i^* - m_j^*\| + \|m_j^*\|$, Eq. (23) certainly holds if either $\|m_i^*\|$ or $\|m_j^*\|$ is upper bounded. Otherwise, if both $\|m_i^*\|$ and $\|m_j^*\|$ increase to infinity, the order of the infinitely large quantity $\|m_i^*\|$ must be lower than or equal to that of $\|m_i^* - m_j^*\|$, which also leads to Eq. (23). Therefore, (ii) holds under the assumptions.

Finally, we turn to (iii) for three cases as follows.

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

$$\begin{aligned}\|m_i^* - m_j^*\|^p(\lambda_{\max}^{i})^{-vq} &= \|m_i^* - m_j^*\|^p(\lambda_{\max}^{i})^{-vp} \\ &= (\eta_i(m_j^*))^{-p} = O(\eta^{-p}(\Phi^*)) = O(\eta^{-p\vee q}(\Phi^*)).\end{aligned}$$

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

$$\|m_i^* - m_j^*\|^p(\lambda_{\max}^{i})^{-vq} \leqslant O(\eta^{-p}(\Phi^*)) = O(\eta^{-p\vee q}(\Phi^*)).$$

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

$$\|m_i^* - m_j^*\|^p(\lambda_{\max}^{i})^{-vq} \leqslant O(\eta^{-q}(\Phi^*)) = O(\eta^{-p\vee q}(\Phi^*)).$$

Summing up the results on the three cases, we have

$$\|m_i^* - m_j^*\|^p(\lambda_{\max}^{i})^{-vq} \leqslant O(\eta^{-p\vee q}(\Phi^*)). \qquad \square$$

Proof of Lemma 2. We begin to prove (i). According to the norm theory, we have

$$\|\Pi_i\| = \|E_{\phi_i^*}((t_i(X) - \phi_i^*)(t_i(X) - \phi_i^*)^{\mathrm{T}})\| \leqslant E_{\phi_i^*}(\|t_i(X) - \phi_i^*\|^2). \tag{37}$$

Since $t_i(x)$ is assumed to be a polynomial of $x_1, x_2, \ldots, x_n$, i.e., the components of x, we transform it into the following expression:

$$t_i(x) = P_0 + P_1x + P_2x^2 + \cdots + P_kx^k,$$

where $k \geqslant 0$, each P_i is a $d_i \times n^i$ matrix, and x^i is a product vector containing all the product terms $x_{j_1}x_{j_2}\cdots x_{j_i}$ as its components, where each x_{j_p} comes from $x_1, x_2, \ldots, x_n$. It can be easily verified that $\|x^i\| \leqslant \sqrt{n}\|x\|^i$ for $i = 0, 1, \ldots, k$.

Based on the above expression, we have

$$\begin{aligned} t_i(x) &= t_i(x - m_i^* + m_i^*) \\ &= P'_0 + P'_1(x - m_i^*) + P'_2(x - m_i^*)^2 + \cdots + P'_k(x - m_i^*)^k, \end{aligned} \tag{38}$$

where each P'_i is still a $d_i \times n^i$ matrix, but its elements are polynomials of $m_{i1}^*, \ldots, m_{in}^*$. We then have

$$\phi_i^* = E_{\phi_i^*}(t_i(X)) = P'_0 + E_{\phi_i^*}(P'_1(X - m_i^*)) + \cdots + E_{\phi_i^*}(P'_k(X - m^*)^k) \tag{39}$$

with $E_{\phi_i^*}(P'_1(X - m_i^*)) = P'_1 E_{\phi_i^*}(X - m_i^*) = 0$, and

$$t_i(X) - \phi_i^* = \sum_{j=1}^{k} [P'_j(X - m_i^*)^j - E_{\phi_i^*}(P'_j(X - m_i^*)^j)].$$

Now, we have

$$\begin{aligned} E_{\phi_i^*}(\|t_i(X) - \phi_i^*\|^2) &= E_{\phi_i^*}(\|(t_i(X) - \phi_i^*)^{\mathrm{T}}(t_i(X) - \phi_i^*)\|) \\ &= E_{\phi_i^*}\left(\left\| \sum_{j_1=1, j_2=1}^{k} [P'_{j_1}(X - m_i^*)^{j_1} - E_{\phi_i^*}(P'_{j_1}(X - m_i^*)^{j_1})]^{\mathrm{T}} \right.\right. \\ &\quad \left.\left. \times [P'_{j_2}(X - m_i^*)^{j_2} - E_{\phi_i^*}(P'_{j_2}(X - m_i^*)^{j_2})] \right\|\right) \\ &\leqslant \sum_{j_1=1, j_2=1}^{k} E_{\phi_i^*}(\|[P'_{j_1}(X - m_i^*)^{j_1} - E(P'_{j_1}(X - m_i^*)^{j_1})]^{\mathrm{T}}\| \\ &\quad \times \|[P'_{j_2}(X - m_i^*)^{j_2} - E_{\phi_i^*}(P'_{j_2}(X - m_i^*)^{j_2})]\|) \\ &\leqslant \sum_{j_1=1, j_2=1}^{k} E^{1/2}_{\phi_i^*}(\|P'_{j_1}(X - m_i^*)^{j_1} - E_{\phi_i^*}(P'_{j_1}(X - m_i^*)^{j_1})\|^2) \\ &\quad \times E^{1/2}_{\phi_i^*}\|P'_{j_2}(X - m_i^*)^{j_2} - E_{\phi_i^*}(P'_{j_2}(X - m_i^*)^{j_2}|\phi_i^*)\|^2). \end{aligned} \tag{40}$$

Particularly, we have

$$\begin{aligned} &E_{\phi_i^*}(\|P'_{j_1}(X - m_i^*)^{j_1} - E_{\phi_i^*}(P'_{j_1}(X - m_i^*)^{j_1})\|^2) \\ &\quad = E_{\phi_i^*}(\|P'_{j_1}(X - m_i^*)^{j_1}\|^2) - \|E_{\phi_i^*}(\|P'_{j_1}(X - m_i^*)^{j_1})\|^2 \\ &\quad \leqslant E_{\phi_i^*}(\|P'_{j_1}(X - m_i^*)^{j_1}\|^2) \leqslant \sqrt{n}E_{\phi_i^*}(\|P'_{j_1}\|^2\|X - m_i^*\|^{2j_1}) \\ &\quad = \sqrt{n}\|P'_{j_1}\|^2 E_{\phi_i^*}(\|X - m_i^*\|^{2j_1}). \end{aligned} \tag{41}$$

Because $P_i(x|\phi_i^*) \leqslant U_i(x|\phi_i^*)$, we further have

$$E_{\phi_i^*}(\|X - m_i^*\|^{2j_1}) \leqslant \int \|x - m_i^*\|^{2j_1} U_i(x|\phi_i^*)\,\mathrm{d}x = \int \|y\|^{2j_1} w(y + m_i^*)(\lambda_{\max}^i)^{-c_1}\,\mathrm{e}^{-\rho(1/(\lambda_{\max}^i)^{\nu c_2})\|y\|^{c_2}}\,\mathrm{d}y, \tag{42}$$

where we take the transformation $y = x - m_i^*$. Since $w(x)$ is a positive polynomial, we certainly have

$$w(y + m_i^*) \leqslant w_0 + w_1\|y\| + \cdots + w_{k'}\|y\|^{k'}, \tag{43}$$

where k' is a positive integer, and $w_0, w_1, \ldots, w_{k'}$ are a group of positive polynomials of $\|m_i^*\|$, i.e.,

$$w_i = w_0^i + w_1^i\|m_i^*\| + \cdots + w_{c_i}^i\|m_i^*\|^{c_i} \quad \text{for } i = 0, 1, \ldots, k', \tag{44}$$

where $w_0^i, w_1^i, \ldots, w_{c_i}^i$ are nonnegative numbers, and $c_0, \ldots, c_{k'}$ are nonnegative integers. By Lemma 1, we further have

$$w_i \leqslant v_0^i + v_1^i\|m_i^* - m_j^*\| + \cdots + v_{c_i}^i\|m_i^* - m_j^*\|^{c_i} \quad \text{for } i = 0, 1, \ldots, k', \tag{45}$$

where $v_0^i, v_1^i, \ldots, v_{c_i}^i$ are nonnegative numbers. Take the upper bound of $w(y + m_i^*)$ into the inequality Eq. (42), we have

$$E_{\phi_i^*}(\|X - m_i^*\|^{2j_1}) \leqslant \sum_{l=0}^{k'} w_l(\lambda_{\max}^i)^{-c_1} \int \|y\|^{2j_1+l}\,\mathrm{e}^{-\rho(1/(\lambda_{\max}^i)^{\nu c_2})\|y\|^{c_2}}\,\mathrm{d}y = \sum_{l=0}^{k'} w_l(\lambda_{\max}^i)^{-c_1+\nu(2j_1+l+1)} \int \|u\|^{2j_1+l}\,\mathrm{e}^{-\rho\|u\|^{c_2}}\,\mathrm{d}u,$$

where we take the transformation $u = y/(\lambda_{\max}^i)^{\nu}$. Clearly, $\int \|u\|^{2j_1+l}\,\mathrm{e}^{-\rho\|u\|^{c_2}}\,\mathrm{d}u$ is finite and upper bounded if j_1 is upper bounded. Since $\lambda_{\max}^i$ is upper bounded, we have that $E_{\phi_i^*}(\|X - m_i^*\|^{2j_1})$ is upper bounded by a positive polynomial of $\|m_i^* - m_j^*\|$.

Moreover, as each element of P'_{j_1} is a polynomial of $m_{i1}^*, \ldots, m_{in}^*$, $\|P'_{j_1}\|$ is upper bounded by a positive polynomial of $\|m_i^*\|$. Therefore, $E_{\phi_i^*}(\|P'_{j_1}(X - m_i^*)^{j_1}\|^2)$ is upper bounded by a positive polynomial of $\|m_i^* - m_j^*\|$.

As a result, $E_{\phi_i^*}(\|P'_{j_1}(X - m_i^*)^{j_1} - E_{\phi_i^*}(P'_{j_1}(X - m_i^*)^{j_1})\|^2)$ is upper bounded by a positive polynomial of $\|m_i^* - m_j^*\|$. Since $\|m_i^* - m_j^*\| \geqslant T'$, we further have that

$$E_{\phi_i^*}(\|P'_{j_1}(X - m_i^*)^{j_1} - E_{\phi_i^*}(P'_{j_1}(X - m_i^*)^{j_1})\|^2) \leqslant \psi_{j_1}\|m_i^* - m_j^*\|^{p_{j_1}}, \tag{46}$$

where ψ_{j_1} and p_{j_1} is some positive numbers.

Take Eq. (46) into Eq. (40), we have

$$E_{\phi_i^*}(\|t_i(X) - \phi_i^*\|^2) \leqslant \psi\|m_i^* - m_j^*\|^p, \tag{47}$$

where ψ and p are positive numbers. Therefore, according to Eq. (37), (i) holds under the assumptions.

As to (ii), when $j \neq i$, we let $\phi'_j = E_{\phi^*_j}(t_i(X))$ and have

$$\begin{aligned} E_{\phi^*_j}(\|t_i(X) - \phi^*_i\|^2) &\leqslant E_{\phi^*_j}((\|t_i(X) - \phi'_j\| + \|\phi'_j - \phi^*_i\|)^2) \\ &= E_{\phi^*_j}(\|t_i(X) - \phi'_j\|^2 + 2\|t_i(X) - \phi'_j\|\|\phi'_j - \phi^*_i\| + \|\phi'_j - \phi^*_i\|^2) \\ &\leqslant E_{\phi^*_j}(2\|t_i(X) - \phi'_j\|^2 + 2\|\phi'_j - \phi^*_i\|^2) \\ &= 2E_{\phi^*_j}(\|t_i(X) - \phi'_j\|^2) + 2\|\phi^*_i - \phi'_j\|^2. \end{aligned} \tag{48}$$

In the same way as above, we can prove that

$$E_{\phi^*_j}(\|t_i(X) - \phi'_j\|^2) \leqslant \psi_1 \|m^*_i - m^*_j\|^{p_1}, \tag{49}$$

where ψ_1 and p_1 are positive numbers. Moreover, we certainly have

$$\|\phi^*_i - \phi'_j\| \leqslant \|\phi^*_i\| + \|\phi'_j\|.$$

By Eq. (38), in a similar way we can prove that both $\|\phi^*_i\|$ and $\|\phi'_j\|$ are upper bounded by $\psi_2\|m^*_i - m^*_j\|^{p_2}$, where ψ_2 and p_2 are some positive numbers. Therefore, by Eq. (48), $E_{\phi^*_j}(\|t_i(X) - \phi^*_i\|^2)$ is upper bounded by a positive polynomial of $\|m^*_i - m^*_j\|$. Since $\|m^*_i - m^*_j\| \geqslant T'$, we have

$$E_{\phi^*_j}(\|t_i(X) - \phi^*_i\|^2) \leqslant \psi_j \|m^*_i - m^*_j\|^{p_j}, \quad j \neq i, \tag{50}$$

where ψ_j and p_j are positive numbers.

By Eqs. (47) and (50), we have

$$E(\|t_i(X) - \phi^*_i\|^2) = \sum_{j=1}^{K} \alpha^*_j E_{\phi^*_j}(\|t_i(X) - \phi^*_i\|^2) \leqslant \upsilon M_i^q(\Phi^*),$$

where $M_i(\Phi^*) = \max_{j \neq i} \|m^*_i - m^*_j\|$, υ and q are positive numbers. □

Proof of Lemma 3. We first prove that

$$f(\eta) = o(\eta^p),$$

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

We consider the mixture of K densities from the bell sheltered exponential families of the parameter Φ^* under the relation $\eta(\Phi^*) = \eta$. When $i \neq j$, for a small enough η, there is certainly a point m^*_{ij} on the line between m^*_i and m^*_j such that

$$\alpha^*_i P_i(m^*_{ij}|\phi^*_i) = \alpha^*_j P_j(m^*_{ij}|\phi^*_j).$$

We further define

$$\begin{aligned} E_i &= \{x : \alpha^*_i P_i(x|\phi^*_i) \geqslant \alpha^*_j P_j(x|\phi^*_j)\}, \\ E_j &= \{x : \alpha^*_j P_j(x|\phi^*_j) > \alpha^*_i P_i(x|\phi^*_i)\}. \end{aligned}$$

As $\eta(\Phi^*)$ tends to zero, $(\lambda^i_{\max})^v/(\|m^*_i - m^*_j\|)$ and $(\lambda^j_{\max})^v/(\|m^*_i - m^*_j\|)$ are the same order infinitesimals. Moreover, $\kappa(\Sigma^*_i)$ and $\kappa(\Sigma^*_j)$ are both upper bounded. Thus, there certainly exists a neighborhood (i.e., a hypersphere) 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. Since $\kappa(\Sigma^*_i)$ and $\kappa(\Sigma^*_j)$ are always

upper bounded, 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

$$r_i \geqslant b_i\|m_i^* - m_j^*\| \quad \text{and} \quad r_j \geqslant b_j\|m_i^* - m_j^*\|.$$

We further define

$$\mathscr{D}_i = \mathscr{N}_{r_i}^c(m_i^*) = \{x\colon \|x - m_i^*\| \geqslant r_i\},$$
$$\mathscr{D}_j = \mathscr{N}_{r_j}^c(m_j^*) = \{x\colon \|x - m_j^*\| \geqslant r_j\}$$

and thus

$$E_i \subset D_j, \qquad E_j \subset D_i.$$

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

$$\begin{aligned} e_{ij}(\Phi^*) &= \int h_i(x)h_j(x)P(x|\Phi^*)\,\mathrm{d}\mu \\ &= \int_{E_i} h_i(x)h_j(x)P(x|\Phi^*)\,\mathrm{d}\mu + \int_{E_j} h_i(x)h_j(x)P(x|\Phi^*)\,\mathrm{d}\mu \\ &\leqslant \int_{\mathscr{D}_j} h_i(x)h_j(x)P(x|\Phi^*)\,\mathrm{d}\mu + \int_{\mathscr{D}_i} h_i(x)h_j(x)P(x|\Phi^*)\,\mathrm{d}\mu \\ &\leqslant \int_{\mathscr{D}_j} h_j(x)P(x|\Phi^*)\,\mathrm{d}\mu + \int_{\mathscr{D}_i} h_i(x)P(x|\Phi^*)\,\mathrm{d}\mu \\ &= \alpha_j^* \int_{\mathscr{D}_j} P_j(x|\phi_j^*)\,\mathrm{d}\mu + \alpha_i^* \int_{\mathscr{D}_i} P_i(x|\phi_i^*)\,\mathrm{d}\mu. \end{aligned}$$

We now consider $\int_{\mathscr{D}_i} P_i(x|\phi_i^*)\,\mathrm{d}\mu$. Since $r_i \geqslant b_i\|m_i^* - m_j^*\|$,

$$\int_{\mathscr{D}_i} P_i(x|\phi_i^*)\,\mathrm{d}\mu \leqslant \int_{\|x-m_i^*\| \leqslant b_i\|m_i^*-m_j^*\|} P_i(x|\phi_i^*)\,\mathrm{d}\mu.$$

By the transformation $y = (x - m_i^*)/\|m_i^* - m_j^*\|$, we have

$$\begin{aligned} &\int_{\mathscr{D}_i} P_i(x|\phi_i^*)\,\mathrm{d}\mu \\ &\quad\leqslant \int_{\|y\| \leqslant b_i} w(\|m_i^* - m_j^*\|y + m_i^*)(\lambda_{\max}^i)^{-c_1} \mathrm{e}^{-\rho(\|m_i^*-m_j^*\|^{c_2})/(\lambda_{\max})^{vc_2}\|y\|^{c_2}} \|m_i^* - m_j^*\|\,\mathrm{d}\mu' \\ &\quad= \int_{\|y\| \leqslant b_i} \|m_i^* - m_j^*\| w(\|m_i^* - m_j^*\|y + m_i^*)(\lambda_{\max}^i)^{-c_1} \\ &\qquad\times \mathrm{e}^{-\rho(\|m_i^*-m_j^*\|^{c_2})/(\lambda_{\max})^{vc_2}\|y\|^{c_2}}\,\mathrm{d}\mu', \end{aligned} \tag{51}$$

where μ' is the transformed measure from μ by the transformation.

Since each coefficient in the polynomial function $w(\|m_i^* - m_j^*\|y + m_i^*)$ is a polynomial function of m_i^* multiplied by a positive order power of $\|m_i^* - m_j^*\|$, there certainly exists a positive number q such that the coefficients in the polynomial function $\|m_i^* - m_j^*\|^{-q} w(\|m_i^* - m_j^*\|y + m_i^*)$ are upper bounded as $\eta(\Phi^*) \to 0$. So

$\|m_i^* - m_j^*\|^{-q} w(\|m_i^* - m_j^*\| y + m_i^*)$ can be upper bounded by another positive polynomial function of y with constant coefficients. Moreover, by Lemma 1, we have

$$\|m_i^* - m_j^*\|^{1+q} (\lambda_{\max}^i)^{-c_1} \leqslant O(\eta^{-c_1'}),$$
$$\|m_i^* - m_j^*\|^{c_2} (\lambda_{\max}^i)^{-vc_2} \geqslant O(\eta^{-c_2}),$$

where $c_1' = (q+1) \vee (c_1/v)$.

According to these results, we have from Eq. (51) that

$$\int_{\mathscr{D}_i} P_i(x|\phi_i^*)\,\mathrm{d}\mu \leqslant \int_{\mathscr{B}_i} \frac{1}{\eta^{c_1'}(\Phi^*)} w_1(y)\,\mathrm{e}^{-\rho'(1/\eta^{c_2}(\Phi^*))\|y\|^{c_2}}\,\mathrm{d}\mu' = \int_{\mathscr{B}_i} \frac{1}{\eta^{c_1'}} w_1(y)\,\mathrm{e}^{-\rho'(1/\eta^{c_2})\|y\|^{c_2}}\,\mathrm{d}\mu', \tag{52}$$

where $\mathscr{B}_i = \{y : \|y\| \geqslant b_i\}$, ρ' is another positive number, and $w_1(y)$ is a positive polynomial function of y with constant coefficients.

Furthermore, we let

$$F_i(\eta) = \int_{\mathscr{B}_i} P(y|\eta)\,\mathrm{d}y, \quad P(y|\eta) = \frac{1}{\eta^{c_1'}} w_1(y)\,\mathrm{e}^{-\rho'(1/\eta^{c_2})\|y\|^{c_2}}$$

and consider the limit of $F_i(\eta)/\eta^p$ as η tends to zero.

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

$$\lim_{\eta\to 0} \frac{P(y|\eta)}{\eta^p} = w_1(y) \lim_{\eta\to 0} \frac{1}{\eta^{(c_1'+p)}}\,\mathrm{e}^{-\rho'(1/\eta^{c_2})\|y\|^{c_2}} = w_1(y) \lim_{\zeta=\frac{1}{\eta}\to\infty} \frac{\zeta^{(c_1'+p)}}{\mathrm{e}^{\zeta^{c_2}\rho'\|y\|^{c_2}}} = 0,$$

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

$$\lim_{\eta\to 0} \frac{F_i(\eta)}{\eta^p} = \lim_{\eta\to 0} \int_{\mathscr{B}_i} \frac{P(y|\eta)}{\eta^p}\,\mathrm{d}\mu' = \int_{\mathscr{B}_i} \lim_{\eta\to 0} \frac{P(y|\eta)}{\eta^p}\,\mathrm{d}\mu' = 0$$

and thus $F_i(\eta) = o(\eta^p)$. It further follows from Eq. (52) that

$$\sup_{\eta(\Phi^*)=\eta} \int_{\mathscr{D}_i} P_i(x|\phi_i^*)\,\mathrm{d}\mu = o(\eta^p). \tag{53}$$

Similarly, we can also prove:

$$\sup_{\eta(\Phi^*)=\eta} \int_{\mathscr{D}_j} P_j(x|\phi_j^*)\,\mathrm{d}\mu = o(\eta^p).$$

As a result, we have

$$\begin{aligned} f_{ij}(\eta) &= \sup_{\eta(\Phi^*)=\eta} e_{ij}(\Phi^*) \\ &\leqslant \sup_{\eta(\Phi^*)=\eta} \left(\alpha_j^* \int_{\mathscr{D}_j} P_j(x|\phi_j^*)\,\mathrm{d}\mu + \alpha_i^* \int_{\mathscr{D}_i} P_i(x|\phi_i^*)\,\mathrm{d}\mu \right) \\ &\leqslant \sup_{\eta(\Phi^*)=\eta} \int_{\mathscr{D}_j} P_j(x|\phi_j^*)\,\mathrm{d}x + \sup_{\eta(\Phi^*)=\eta} \int_{\mathscr{D}_i} P_i(x|\phi_i^*)\,\mathrm{d}\mu) \\ &= o(\eta^p). \end{aligned}$$

Thus, we have

$$f(\eta) \leqslant \max_{ij} f_{ij}(\eta) = o(\eta^p). \tag{54}$$

Moreover, because

$$\lim_{\eta\to 0} \frac{f^{\varepsilon}(\eta)}{\eta^p} = \lim_{\eta\to 0} \left(\frac{f(\eta)}{\eta^{\frac{p}{\varepsilon}}} \right)^{\varepsilon} = 0,$$

we finally have $f^{\varepsilon}(\eta) = o(\eta^p)$ and thus $f^{\varepsilon}(\eta(\Phi^*)) = o(\eta^p(\Phi^*))$. □

References

[1] O.E. Barndorf-Nielsen, Information and Exponential Families in Statistical Theory, Wiley, New York, 1978.

[2] B. Delyon, M. Lavielle, E. Moulines, Convergence of a stochastic approximation version of the EM algorithm, Ann. Statist. 27 (1999) 94–128.

[3] 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.

[4] S.C. Horng, Examples of sublinear convergence of the EM algorithm, in: Proceeding of the Statistical Computing Section, American Statistical Association, 1987, pp. 266–271.

[5] M. Jamshidian, R.I. Jennrich, Conjugate gradient acceleration of the EM algorithm, J. Am. Statist. Assoc. 88 (1993) 221–228.

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

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

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

[9] N. Laird, N. Lange, D. Stram, Maximizing likelihood computations with repeated measures: application of the EM algorithm, J. Am. Statist. Assoc. 82 (1987) 97–105.

[10] K. Lange, A quasi-Newtonian acceleration of the EM algorithm, Statist. Sin. 5 (1995) 1–18.

[11] E.L. Lehmann, Theory of Point Estimation, Wiley, New York, 1985.

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

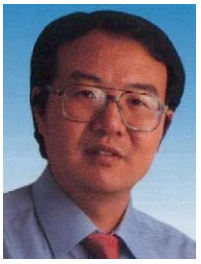
[13] C. Liu, D.B. Rubin, Y. Wu, Parameter expansion to accelerate EM: the PX-EM algorithm, Biometrika 85 (1998) 755–770.

[14] J. Ma, L. Xu, M.I. Jordan, Asymptotic convergence rate of the EM algorithm for Gaussian mixtures, Neural Comput. 12 (2000) 2881–2907.

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

[16] 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.
[17] L.R. Rabiner, A tutorial on Hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–286.
[18] R.A. Redner, H.F. Walker, Mixture densities, maximum likelihood, and the EM algorithm, SIAM R. 26 (1984) 195–239.
[19] C.F.J. Wu, On the convergence properties of the EM algorithm, Ann. Statist. 11 (1983) 95–103.
[20] L. Xu, M.I. Jordan, On convergence properties of the EM algorithm for Gaussian mixtures, Neural Comput. 8 (1996) 129–151.

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 the Department of Mathematics, Shantou University. From December 1999, he worked as a full professor at the Institute of Mathematic, Shantou University. In September 2001, he was transferred to the Department of Information Science at the School of Mathematical Sciences, Peking University. During 1995 and 2004, he also visited and studied several times at Department of Computer Science & Engineering, the Chinese University of Hong Kong as a Research Associate or Fellow. He has published over 50 academic papers on neural networks, pattern recognition, artificial intelligence, and information theory.



Lei Xu (IEEE Fellow) is a chair professor of Computer Science & Engineering, Chinese University of Hong Kong. He completed his Ph.D. thesis at Tsinghua University by the end of 1986, then joined Department of Mathematics, Peking University in 1987 first as a postdoc and then was exceptionally promoted to associate professor in 1988. During 1989–93, he worked at several universities in Finland, Canada and USA, including Harvard and MIT. He joined CUHK in 1993 as senior lecturer, became professor in 1996 and took the current chair professor in 2002. Prof. Xu has served or is serving as associate editor for several journals, including Neural Networks, IEEE Transactions on Neural Networks, as a governor of International Neural Network Society(01-03), the chair of Computational Finance Technical Committee of IEEE Computational Intelligence Society(01-03), and a past president of Asian-Pacific Neural Networks Assembly. Prof. Xu has published nearly 100 academic journal papers on statistical learning, pattern recognition, neurocomputing, and computational finance, with a number of them well cited in the fields. He has given a number of keynote/plenary/invited/tutorial in major international conferences in the fields, and also served as program committee chair and general chair for a number of international conferences. He has received several Chinese national prestigious academic awards (including 1994 Chinese National Nature Science Award) and international awards (including 1995 INNS Leadership Award). Prof. Xu is an IEEE Fellow and a Fellow of International Association for Pattern Recognition, and a member of European Academy of Sciences.