Pattern Recognition 45 (2012) 3950–3961

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

A robust EM clustering algorithm for Gaussian mixture models

Miin-Shen Yang*, Chien-Yo Lai, Chih-Ying Lin

Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan

ARTICLE INFO

Article history:
Received 23 November 2011
Received in revised form
18 March 2012
Accepted 24 April 2012
Available online 14 May 2012

Keywords:
Cluster analysis
EM algorithm
Gaussian mixture model
Robust EM
Initialization
Number of clusters

ABSTRACT

Clustering is a useful tool for finding structure in a data set. The mixture likelihood approach to clustering is a popular clustering method, in which the EM algorithm is the most used method. However, the EM algorithm for Gaussian mixture models is quite sensitive to initial values and the number of its components needs to be given a priori. To resolve these drawbacks of the EM, we develop a robust EM clustering algorithm for Gaussian mixture models, first creating a new way to solve these initialization problems. We then construct a schema to automatically obtain an optimal number of clusters. Therefore, the proposed robust EM algorithm is robust to initialization and also different cluster volumes with automatically obtaining an optimal number of clusters. Some experimental examples are used to compare our robust EM algorithm with existing clustering methods. The results demonstrate the superiority and usefulness of our proposed method.

1. Introduction

Data analysis is a science for analyzing data in real world, and cluster analysis is a useful tool for data analysis. Cluster analysis is a method for finding clusters within a data set characterized by the greatest similarity within the same cluster and the greatest dissimilarity between different clusters. Hierarchical clustering was the earliest clustering method used by biologists and social scientists, whereas cluster analysis became a branch of statistical multivariate analysis. Many theories and methods for cluster analysis have been presented in the literature [1–3]. In general, learning and recognition mostly start from clustering, so that cluster analysis becomes a type of unsupervised learning in pattern recognition and has been widely applied in various areas [4].

From the statistical point of view, clustering methods may be divided into probability model-based approaches and nonparametric approaches. The probability model-based approach assumes that the data set follows a mixture model of probability distributions so that a mixture likelihood approach to clustering may be used [2]. For a mixture model, the expectation and maximization (EM) algorithm [5] is commonly used. For a nonparametric approach, clustering methods may be based on an objective function of similarity or dissimilarity measures, and these can be divided into hierarchical and partitional methods. A hierarchical clustering method is a procedure for transforming a data set into a diagram, known as a *dendrogram*, based on the similarity or dissimilarity matrix of the data set. Most partitional methods suppose that the data set can be represented by finite cluster prototypes with their own objective functions. Therefore, defining the dissimilarity (or distance) between a point and a cluster prototype is essential for partition methods. The most popular partition methods with cluster prototypes are *k*-means [6,7], trimmed *k*-means [8,9], fuzzy c-means (FCM) [10,11], and mean shift [12,13].

In this paper we focus on clustering based on probability models, and in particular, we propose a robust type of EM algorithm for Gaussian mixture models. We know that the EM algorithm is quite sensitive to initial values, in which the number of components needs to be given a priori. In this paper we present a robust EM clustering algorithm which will be robust to initials and different cluster volumes with automatically obtaining an optimal number of clusters. Although some authors have considered the initial problems for the EM algorithm [14,15] and some have considered estimation of the number of components [15,16], there has been less consideration about robustness to initial values associated with the number of components for the EM algorithm. Since this robustness property is very important for the EM, we present a new means of solving these initial problems by automatically finding an optimal number of components. We first propose a new objective function based on mixture distributions and then create new update equations for the EM algorithm. We also construct a learning schema to automatically obtain an optimal number of components.

The rest of the paper is organized as follows. In Section 2, we briefly review the EM algorithm. In Section 3, we propose a robust

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

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.04.031

EM clustering algorithm. In Section 4, we use our algorithm with some artificial datasets and real datasets to demonstrate that this algorithm is effective in Gaussian mixture models. Finally, we state conclusions in Section 5.

2. The EM clustering algorithm

Let the data set $\{X_1, X_2,\ldots,X_n\}$ be a random sample of size n from the d-variate mixture model

$$f(x;\alpha,\theta)=\sum_{k=1}^{c}\alpha_k f(x;\theta_k) \tag{1}$$

where $\alpha_k>0$ denotes mixing proportions with the constraint $\sum_{k=1}^{c}\alpha_k=1$ and $f(x;\theta_k)$ denotes the density of x from kth class with corresponding parameters θ_k. Let $Z=\{Z_1, Z_2,\ldots,Z_n\}$ be the missing data in which $Z_i\in\{1, 2,\ldots,c\}$. If $Z_i=k$, it means that the ith data point belongs to the kth class. Thus, the joint pdf of the complete data $\{X_1, X_2,\ldots, X_n, Z_1, Z_2,\ldots,Z_n\}$ becomes

$$f(x_1,\ldots,x_n,z_1,\ldots,z_n;\alpha,\theta)=\mathop{\Pi}_{i=1}^{n}\mathop{\Pi}_{k=1}^{c}[\alpha_k f(x_i;\theta_k)]^{z_{ki}} \tag{2}$$

where $z_{ki}=\begin{cases}1, & if \quad Z_i=k\\ 0, & if \quad Z_i\neq k\end{cases}$. The log likelihood function is obtained as follows:

$$L(\alpha,\theta;x_1,\ldots,x_n,z_1,\ldots,z_n)=\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ki}\ln[\alpha_k f(x_i;\theta_k)] \tag{3}$$

E-step: Since the latent variables z_{ki} are unknown, according to Dempster et al. [5], the conditional expected value $E(Z_{ki}|x_i;\alpha,\theta)$ is substituted for z_{ki}. By Baye's Theorem, we have

$$\hat{z}_{ki}=E(Z_{ki}|x_i;\alpha,\theta)=\frac{\alpha_k f(x_i;\theta_k)}{\sum_{s=1}^{c}\alpha_s f(x_i;\theta_s)} \tag{4}$$

M-step: Under the constraint $\sum_{k=1}^{c}\alpha_k=1$, to maximize

$$\tilde{L}(\alpha,\theta;x_1,\ldots,x_n)=\sum_{i=1}^{n}\sum_{k=1}^{c}\hat{z}_{ki}\ln[\alpha_k f(x_i;\theta_k)] \tag{5}$$

We can obtain the updated equation for mixing proportions with

$$\alpha_k=\frac{\sum_{i=1}^{n}\hat{z}_{ki}}{n} \tag{6}$$

We now consider the d-variate Gaussian mixture model

$$\begin{aligned}f(x;\alpha,\theta)&=\sum_{k=1}^{c}\alpha_k f(x;\theta_k)\\&=\sum_{k=1}^{c}\alpha_k(2\pi)^{-(d/2)}|\Sigma_k|^{-(1/2)}e^{-(1/2)(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)}\end{aligned} \tag{7}$$

The parameter θ_k consists of a mean vector μ_k and a covariance matrix Σ_k. Then the update equations for those parameters are as follows:

$$\mu_k=\frac{\sum_{i=1}^{n}\hat{z}_{ki}x_i}{\sum_{i=1}^{n}\hat{z}_{ki}} \tag{8}$$

$$\Sigma_k=\frac{\sum_{i=1}^{n}\hat{z}_{ki}(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{n}\hat{z}_{ki}} \tag{9}$$

Thus, the EM clustering algorithm can be summarized as follows.

EM clustering algorithm for normal mixtures

```
Step 1: Fix $2\le c\le n$ and fix any $\varepsilon>0$.
    Give initials $\hat{z}^{(0)}=(\hat{z}_1^{(0)},\ldots,\hat{z}_c^{(0)})$ and let $s=1$.
Step 2: Compute $\alpha^{(s)}$ and $\mu^{(s)}$ with $\hat{z}^{(s-1)}$ using (6) and (8).
Step 3: Compute $\sum^{(s)}$ with $\hat{z}^{(s-1)}$ and $\mu^{(s)}$ using (9).
Step 3: Update $\hat{z}^{(s)}$ with $(\alpha^{(s)}, \mu^{(s)}, \sum^{(s)})$ using (4).
Step 4: Compare $\hat{z}^{(s)}$ to $\hat{z}^{(s-1)}$ in a convenient matrix norm $\|\cdot\|$.
    IF $\|\hat{z}^{(s)}-\hat{z}^{(s-1)}\|<\varepsilon$, STOP
    ELSE $s=s+1$ and return to step 2.
```

We mention that the convergence properties of the EM algorithm had been well discussed in Wu [17]. Afterwards, Xu and Jordan [18] considered more convergence properties of the EM algorithm for Gaussian mixtures. Ma et al. [19] considered the convergence rate of the EM algorithm for Gaussian mixtures. Since the EM algorithm is quite sensitive to initialization, in which the cluster numbers need to be given a priori, Figueiredo and Jain [20] proposed an algorithm to deal simultaneously with the number of clusters and also the estimates of parameters for mixture models by using the particular form of a minimum message length (MML) criterion. This criterion is the minimization of the following cost function via EM estimators

$$\begin{aligned}K(\alpha,\theta;x_1,\ldots,x_n)=&\frac{P}{2}\sum_{m:\alpha_m>0}\ln\left(\frac{n\alpha_m}{12}\right)+\frac{c_{nz}}{2}\ln\left(\frac{n}{12}\right)+\frac{c_{nz}(P+1)}{2}\\&-\sum_{i=1}^{n}\ln\left[\sum_{k=1}^{c}\alpha_k f(x_i;\theta_k)\right]\end{aligned} \tag{10}$$

where P is the number of parameters specifying each component and c_{nz} denotes the number of non-zero-probability components. Then the update equation for the proportion is as follows:

$$\alpha_k=\frac{\max\{0,\sum_{i=1}^{n}\hat{z}_{ki}-\frac{P}{2}\}}{\sum_{s=1}^{c}\max\{0,\sum_{i=1}^{n}\hat{z}_{si}-\frac{P}{2}\}} \tag{11}$$

In the d-variate Gaussian mixture model, $\theta_k=(\mu_k,\Sigma_k)$, $P=d+(d(d+1)/2)$ and the update equations of $\hat{z}_{ki}$, μ_k, and Σ_k are the same as the formulas (4), (8) and (9), respectively. For updating parameters, Figueiredo and Jain [20] used the component-wise EM method proposed by Celeux et al. [21], in which the parameters are sequentially updated. The algorithm proposed by Figueiredo and Jain [20] performs first by inputting larger cluster numbers and then by using the formula (11) to eliminate these smaller clusters to reduce the cluster number. After that, they use the criterion (10) to find the clustering that best minimizes the criterion. However, using random initial conditions for the EM in Figueiredo and Jain [20] still has an initialization problem in which using larger initial cluster numbers only makes this initialization problem lighter. We give an example to illustrate it as follows.

Example 1. In this example we use a data set, as shown in Fig. 1(a), generated from a two-component Gaussian mixture distribution with a sample size 800 and the parameters

$$\alpha_1=\alpha_2=0.5,\quad \mu_1=(0\quad 0)^T,\quad \mu_2=(20\quad 0)^T,$$
$$\Sigma_1=\begin{pmatrix}1&0\\0&1\end{pmatrix},\quad \text{and}\quad \Sigma_2=\begin{pmatrix}9&0\\0&9\end{pmatrix}.$$

We use the starting cluster number $c_{initial}=30$ and 100 different random initial conditions for the algorithm of Figueiredo and Jain [20]. Finally we have 78 of 100 with the results of $c^*=2$, as shown in Fig. 1(b), another 11 of 100 with the results of $c^*=3$, and the other 11 of 100 with the results of $c^*>3$. Fig. 1(c) demonstrates an incorrect clustering result of $c^*=3$. In fact, if a data set with a larger cluster number is considered, then the initialization problem for the algorithm of Figueiredo and Jain [20] will become more serious. In next section, we will compare the algorithm in Figueiredo and Jain [20] with our proposed robust EM clustering algorithm.

3. A robust EM clustering algorithm

In this section, we first consider the α_k terms to adjust the EM mixture objective function (3). We know that the proportion α_k can be the probability of one data point belonged to the kth class. Hence, we can use $-\ln\alpha_k$ as the information in the occurrence of one data point belonged to the kth class, and $-\sum_{k=1}^{c}\alpha_k \ln\alpha_k$ is the average of information, which is generally called entropy. When $\alpha_k=1/c$, $\forall k=1,2,\ldots,c$, we say that there is no information about α_k. At this point, we have the entropy achieve the maximum value. Therefore, we first add this term to the original EM objective function. We then use a learning process to estimate α_k by minimizing the entropy to get the most information for α_k. To minimize $-\sum_{k=1}^{c}\alpha_k \ln\alpha_k$ is equivalent to maximizing $\sum_{k=1}^{c}\alpha_k \ln\alpha_k$. For this reason, we use $\sum_{k=1}^{c}\alpha_k \ln\alpha_k$ as a penalty term for the EM mixture objective function. Thus our first proposed EM mixture objective function is to maximize

$$J(\alpha,\theta)=\sum_{i=1}^{n}\sum_{k=1}^{c}\hat{z}_{ki}\ln[\alpha_k f(x_i;\theta_k)]+\beta\sum_{i=1}^{n}\sum_{k=1}^{c}\alpha_k \ln\alpha_k,\quad \beta\geq 0. \tag{12}$$

We know that the EM is a very good algorithm for estimating parameters when good initial conditions are given. In formula (12), the penalty term $\sum_{i=1}^{n}\sum_{k=1}^{c}\alpha_k \ln\alpha_k$ is just used to adjust the EM algorithm in which it is always negative. Hence, at the end of performing this algorithm, we let β be zero to obtain the original EM estimate. The α_k proportions can be derived by maximizing $J(\alpha,\theta)$ with respect to α_k under the constraint $\sum_{k=1}^{c}\alpha_k=1$ (see Appendix) with the following update equation:

$$\alpha_k^{(new)}=\alpha_k^{EM}+\beta\alpha_k^{(old)}(\ln\alpha_k^{(old)}-\sum_{s=1}^{c}\alpha_s^{(old)}\ln\alpha_s^{(old)}) \tag{13}$$

where $\alpha_k^{EM}=\sum_{i=1}^{n}\hat{z}_{ki}/n$.

We should mention that the Eq. (13) created above is important for our proposed robust EM clustering method. In Eq. (13), $\sum_{s=1}^{c}\alpha_s\ln\alpha_s$ is the weighted mean of $\ln\alpha_k$ with the weights $\alpha_1,\ldots,\alpha_c$. For the kth mixing proportion $\alpha_k^{(old)}$, if $\ln\alpha_k^{(old)}$ is less than the weighted mean, then the new mixing proportion $\alpha_k^{(new)}$ will become smaller than the old $\alpha_k^{(old)}$. That is, the smaller proportion will decrease and the bigger proportion will increase in the next iteration and then competition will occur. This situation is like formula (11) used in Figueiredo and Jain [20]. If $\alpha_k\leq 0$ or $\alpha_k<1/n$ for some $1\leq k\leq c^{(old)}$, they are considered to be illegitimate proportions. In this situation, we discard those clusters (or set those proportions to become zero) and then update the cluster number $c^{(old)}$ to

$$c^{(new)}=c^{(old)}-|\alpha_k|\alpha_k<1/n,\quad k=1,\cdots,c^{(old)}| \tag{14}$$

Furthermore, in order to retain the constraints $\sum_{k'=1}^{c^{(new)}}\alpha_{k'}=1$ and $\sum_{k'=1}^{c^{(new)}}\hat{z}_{k'i}=1$, we adjust $\alpha_{k'}$ and $\hat{z}_{k'i}$ by

$$\alpha_{k'}=\frac{\alpha_{k'}}{\sum_{s=1}^{c^{(new)}}\alpha_s} \tag{15}$$

$$\hat{z}_{k'i}=\frac{\hat{z}_{k'i}}{\sum_{s=1}^{c^{(new)}}\hat{z}_{si}} \tag{16}$$

Under our competition schema setting, the algorithm can automatically reduce the number of clusters and also simultaneously get the estimates of parameters. On the other hand, the parameter β can help us control the competition. We discuss the variable β as follows. The plot of the function $f(\alpha)=\alpha\ln\alpha$ is shown in Fig. 2. We can derive that

$$-e^{-1}\leq\alpha_k\ln\alpha_k<0 \tag{17}$$

If $0<\alpha_k\leq 1$, $\forall k=1,2,\ldots,c$, and let

$$E=\sum_{s=1}^{c}\alpha_s\ln\alpha_s<0 \tag{18}$$

then

$$\alpha_k E=\alpha_k\sum_{s=1}^{c}\alpha_s\ln\alpha_s<0 \tag{19}$$

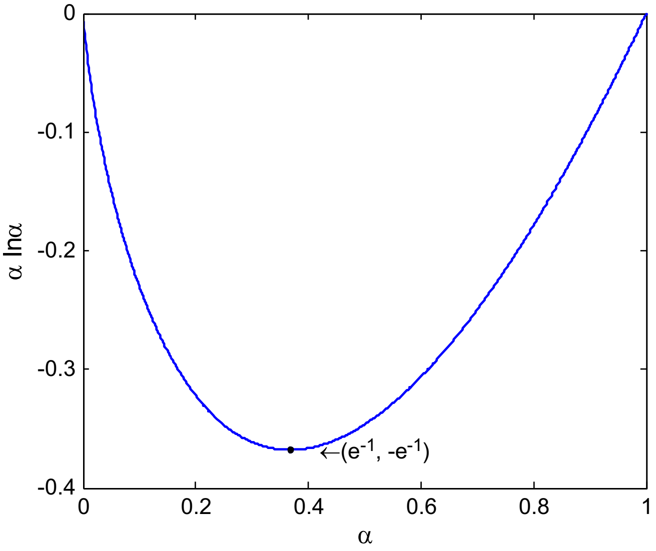

Fig. 2. The plot of function$f(\alpha)=\alpha\ln\alpha$, $0<\alpha\leq 1$.

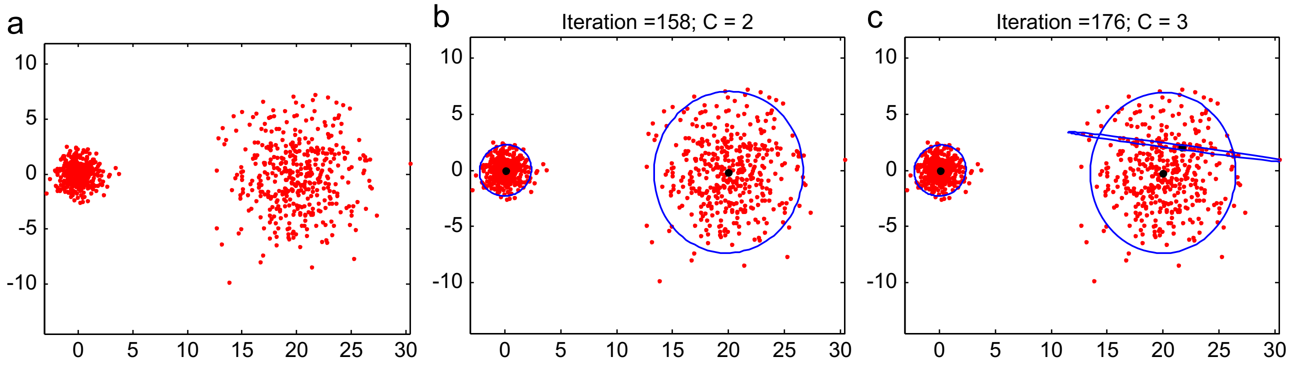

Fig. 1. (a) Data set; (b) clustering results of $c^{*}=2$ and (c) clustering results of $c^{*}=3$.

Using the formula (17) and (19), we can have

$$-e^{-1}\beta < \beta\alpha_k\left(\ln \alpha_k - \sum_{s=1}^{c} \alpha_s \ln \alpha_s\right) < \beta(-\alpha_k E) \tag{20}$$

Under the constraint $\sum_{k=1}^{c} \alpha_k = 1$, and only when $\alpha_k < 1/2$, we can have that $(\ln \alpha_k - \sum_{s=1}^{c} \alpha_s \ln \alpha_s) < 0$. To avoid the situation where all $\alpha_k \le 0$, the left hand of inequality (20) must be larger than $-\max\alpha_k \,|\, \alpha_k < 1/2, \quad k=1,2,\cdots,c$. We now have an elementary condition of β as follows:

$$-e^{-1}\beta > -\max\alpha_k \,|\, \alpha_k < 1/2, \quad k=1,2,\cdots,c$$

$$\beta < \max\alpha_k e \,|\, \alpha_k < 1/2, \quad k=1,2,\cdots,c < e/2. \tag{21}$$

Therefore, to prevent β from being too big, we can use $\beta \in [0, 1]$. Furthermore, if the difference between $\alpha^{(new)}$ and $\alpha^{(old)}$ is small, then β must become large in order to enhance its competition. If the difference between $\alpha^{(new)}$ and $\alpha^{(old)}$ is large, then β will become small to maintain stability. Thus, we define an updated formula for β as

$$\beta = \frac{\sum_{k=1}^{c} \exp{-\eta n\left|\alpha_k^{(new)} - \alpha_k^{(old)}\right|}}{c} \tag{22}$$

where η can be set to be $\min\left\{1, \quad 0.5^{\left\lfloor \frac{d}{2}-1 \right\rfloor}\right\}$, where $\lfloor a \rfloor$ denotes the largest integer that is no more than a. By combining formula (13) and the right hand of inequality (20) to avoid $\alpha_{(1)}^{(new)} = \max_{1 \le k \le c} \alpha_k^{(new)} > 1$, we let $\alpha_{(1)}^{EM} = \max_{1 \le k \le c} \alpha_k^{EM}$, $\alpha_{(1)}^{(old)} = \max_{1 \le k \le c} \alpha_k^{(old)}$, and $E = \sum_{k=1}^{c} \alpha_k^{(old)} \ln \alpha_k^{(old)}$. Then

$$\alpha_{(1)}^{EM} + \beta(-\alpha_{(1)}^{(old)} E) \le 1$$

$$\beta \le (1 - \alpha_{(1)}^{EM})/(-\alpha_{(1)}^{(old)} E) \tag{23}$$

Therefore, we obtain the formula for β as follows:

$$\beta = \min\left\{\frac{\sum_{k=1}^{c} \exp(-\eta n\left|\alpha_k^{(new)} - \alpha_k^{(old)}\right|)}{c}, \frac{(1-\alpha_{(1)}^{EM})}{(-\alpha_{(1)}^{(old)} E)}\right\} \tag{24}$$

Now, let us consider the Gaussian mixture model

$$\begin{aligned} f(x;\alpha,\theta) &= \sum_{k=1}^{c} \alpha_k f(x;\mu_k,\Sigma_k) \\ &= \sum_{k=1}^{c} \alpha_k (2\pi)^{-(d/2)} \left|\Sigma_k\right|^{-(1/2)} e^{-(1/2)(x-\mu_k)' \Sigma_k^{-1} (x-\mu_k)} \end{aligned}$$

The updated equations for parameters μ_k and Σ_k can be derived as follows:

$$\mu_k = \frac{\sum_{i=1}^{n} \hat{z}_{ki} x_i}{\sum_{i=1}^{n} \hat{z}_{ki}} \tag{25}$$

$$\Sigma_k = \frac{\sum_{i=1}^{n} \hat{z}_{ki} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} \hat{z}_{ki}} \tag{26}$$

Because the β can jump at any time, we let $\beta=0$ when the cluster number c is stable. When cluster number c is stable, that means c is no longer decreasing. In our setting, we use all data points as initial means $\mu_k = x_k$, i.e. $c^{initial} = n$, and we use $\alpha_k = 1/c^{initial}$, $\forall k = 1, 2, \ldots, c^{initial}$ as the initial mixing proportions. We also choose a feasible initial covariance matrix. Let

$$\begin{aligned} D_k &= sort\{d_{ki}^2 = \left\|x_i - \mu_k\right\|^2 : d_{ki}^2 > 0, \quad i \ne k, \quad 1 \le i \le n\} \\ &= \{d_{k(1)}^2, d_{k(2)}^2, \ldots, d_{k(n')}^2\} \end{aligned}$$

and

$$\Sigma_k = d^2_{k\left(\left\lceil \sqrt{c^{initial}} \right\rceil\right)} \mathbf{I}_d \tag{27}$$

where $\boldsymbol{I}_d$ is a dxd identity matrix.

When we use a larger cluster number to the EM for Gaussian distribution, the covariance matrix of the cluster with a very small proportion α_k may be close to singular. To avoid this problem, we use a constrain covariance matrix $\tilde{\Sigma}_k$ as follows:

$$\tilde{\Sigma}_k = (1-\gamma)\Sigma_k + \gamma Q \tag{28}$$

where γ is a small positive number and Q is also a diagonal matrix with small positive numbers on its diagonal. In this paper, we use $\gamma = 0.0001$, $Q = d_{\min}^2 \mathbf{I}_d$, $d_{\min}^2 = \min\{d_{ij}^2 : d_{ij}^2 = \left\|x_i - x_j\right\|^2 > 0,\ 1 \le i, j \le n\}$. Thus, the proposed robust EM clustering algorithm can be summarized as follows.

Robust EM clustering algorithm

```
Step 1: Fix $\varepsilon > 0$.
    Give initial $\beta^{(0)}=1$, $c^{(0)}=n$, $\alpha_k^{(0)} = 1/n$, and $\mu^{(0)}=X$.
Step 2: Compute $\Sigma_k^{(0)}$ by (27).
Step 3: Compute $\hat{z}_{ki}^{(0)}$ with $\alpha_k^{(0)}$, $\mu_k^{(0)}$, and $\Sigma_k^{(0)}$ by (4) and set $t=1$.
Step 4: Compute $\mu_k^{(t)}$ with $\hat{z}_{k1}^{(t-1)},\ldots,\hat{z}_{kn}^{(t-1)}$ by (25).
Step 5: Update $\alpha_k^{(t)}$ with $\hat{z}_{k1}^{(t-1)},\ldots,\hat{z}_{kn}^{(t-1)}$ and $\alpha_k^{(t-1)}$ by (13).
Step 6: Compute $\beta^{(t)}$ with $\alpha^{(t)}$ and $\alpha^{(t-1)}$ by (24).
Step 7: Update $c^{(t-1)}$ to $c^{(t)}$ by discard those clusters with
$\alpha_k^{(t)} \le 1/n$ and adjust $\alpha_k^{(t)}$ and $\hat{z}_{ki}^{(t-1)}$ by (15) and (16).
    IF $t \ge 60$ and $c^{(t-60)} - c^{(t)} = 0$, THEN let $\beta^{(t)}=0$.
Step 8: Update $\Sigma_k^{(t)}$ with $\mu_k^{(t)}$ and $\hat{z}_{k1}^{(t-1)},\ldots,\hat{z}_{kn}^{(t-1)}$ by (26) and (28).
Step 9: Update $\hat{z}_{ki}^{(t)}$ with $\alpha_k^{(t)}$, $\mu_k^{(t)}$, and $\Sigma_k^{(t)}$ by (4).
Step 10: Update $\mu_k^{(t+1)}$ with $\hat{z}_{k1}^{(t)},\ldots,\hat{z}_{kn}^{(t)}$ by (25).
Step 11: Compare $\mu^{(t+1)}$ and $\mu^{(t)}$.
    IF $\max_{1 \le k \le c^{(t)}} \left\|\mu_k^{(t+1)} - \mu_k^{(t)}\right\| < \varepsilon$, THEN Stop.
    ELSE $t=t+1$ and return to Step 5.
```

In the Step 1 of the proposed robust EM clustering algorithm, we had assigned all data points as initial values with $c^{(0)}=n$, $\alpha_k^{(0)} = 1/n$, and $\mu^{(0)}=X$. In fact, we can let our robust EM clustering algorithm have the same running way as the method of Figueiredo and Jain [20] that uses some random initial values by only inputting a larger cluster number, but not assigning all data points as initial values. In this case, we may randomly choose $c_{initial}$ data points from the data set that $c_{initial}$ may be less than n with the initial means $\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_{c_{initial}}^{(0)}$. Then the initial proportions will become $\alpha_k^{(0)} = 1/c_{initial}$ for $k=1, 2, \ldots, c_{initial}$. The other steps (i.e. Step 2–11) in our robust EM clustering algorithm keep no change.

We next compare the proposed robust EM clustering algorithm with the EM and the algorithm proposed by Figueiredo and Jain [20] (called the FJ algorithm).

Example 1 (Continued). In this example, we continue Example 1 in Section 2, implementing our robust EM clustering algorithm for the data set of Fig. 1(a). This robust EM clustering algorithm uses all data points as the initial clustering centers, as shown in Fig. 3(a). After the iteration has been implemented once, we can see that the cluster number decreases rapidly from 800 to 587, as shown in Fig. 3(b), then from 587 to 59 after 10 iterations, and so forth. Finally, this data set is successfully grouped in 2 clusters by the robust EM when iteration=41, as shown in Fig. 3(f). In this sense, our robust EM clustering algorithm does not depend on initial cluster centers and can easily obtain correct clustering results. Furthermore, we use the same parameters to generate 100 data sets from the two-component Gaussian mixture distribution and then implement our robust EM clustering algorithm for these 100 data

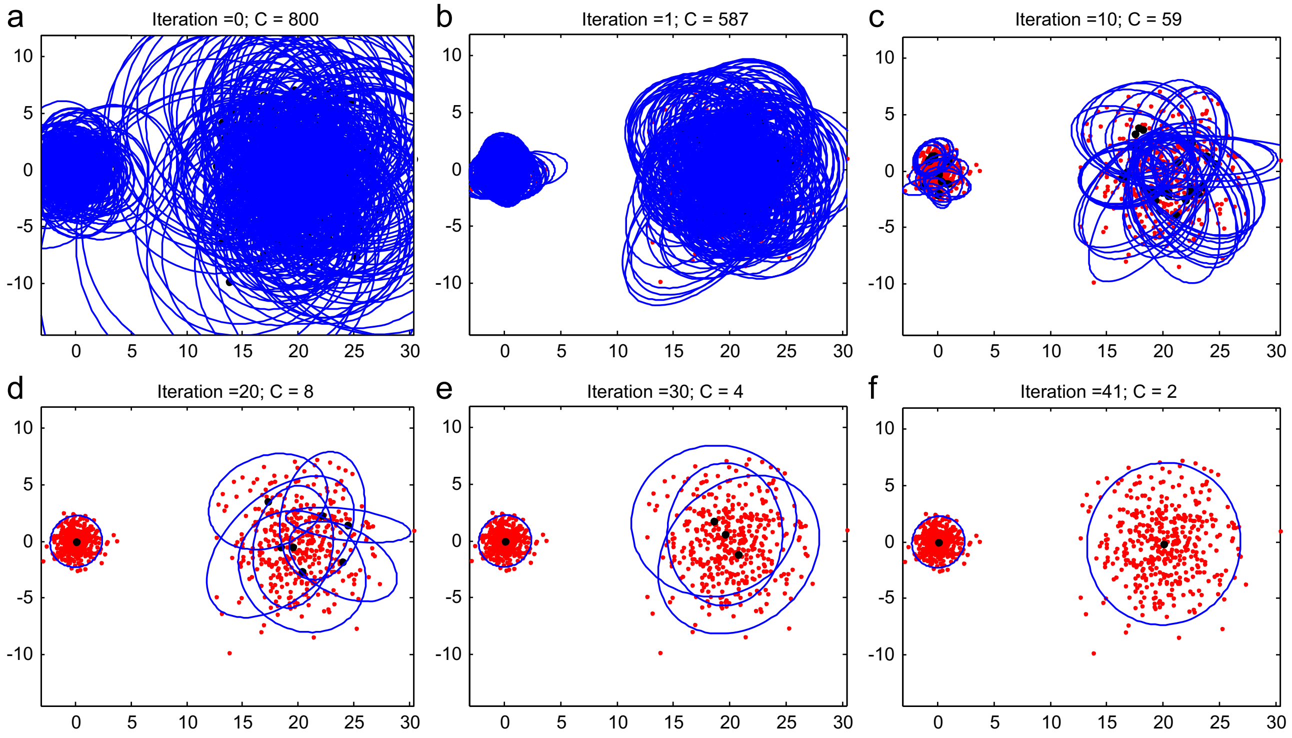

Fig. 3. (a) Initialization of the robust EM; (b)–(e) processes of the robust EM after 1, 10, 20, and 30 iterations and (f) convergent results of the robust EM after 41 iterations.

sets. We find that there are 80 of 100 data sets in which our robust EM clustering algorithm gives correct clustering with $c^*=2$. We recall that there are 78 out of 100 initials with the correct results of $c^*=2$ for the FJ algorithm [20], as shown in Fig. 1(b), so that the FJ algorithm is dependent on initialization.

Furthermore, we also compare our robust EM clustering algorithm with the FJ algorithm under the same initial values by starting cluster number $c_{initial}=30$ and 100 different random initial conditions that had been done in Example 1 of Section 2. For the FJ algorithm, we know that there are 78 of 100 with the results of $c^*=2$ and the average iteration is 209.49. For our robust EM clustering algorithm, there are 94 of 100 with the results of $c^*=2$ and the average iteration is 43.48. We also generate 100 data sets and then use $c_{initial}=30$ and 100 different random initial conditions to each data set. We find that the FJ algorithm obtains the accuracy rate 0.7816 with the results of $c^*=2$. For our robust EM clustering algorithm, the accuracy rate 0.7954 with results of $c^*=2$. Totally, our robust EM clustering algorithm actually presents better than the FJ algorithm.

Example 2. The data set shown in Fig. 4(a) is generated from a two-dimensional, three-component Gaussian mixture distribution with sample size $n=300$ and the parameters

$$\alpha_1=\alpha_2=\alpha_3=1/3, \quad \mu_1=(0 \quad 3)^T, \quad \mu_2=(0 \quad 5)^T,$$

$$\mu_3=(0 \quad 7)^T, \quad \Sigma_1=\Sigma_2=\Sigma_3=\begin{pmatrix}1.2 & 0\\ 0 & 0.01\end{pmatrix}$$

This is a well-separated data set. If the clustering number 3 is given for implementing the EM algorithm with different random initials, then the EM algorithm most has perfect clustering results with $c^*=3$, as shown in Fig. 4(b). If we use the FJ algorithm [20] with random initials and the starting clustering number 20, then it also obtain perfect clustering results with $c^*=3$, as shown in Fig. 4(b). If we use the proposed robust EM, it also obtain perfect clustering results with $c^*=3$ without an initial cluster number, as shown in Fig. 4(b).

Now, we add a two-dimensional Gaussian data set with the sample size 100, and the parameters $\mu_4=(0 \quad 5)^T$ and $\Sigma_4=\begin{pmatrix}0.01 & 0\\ 0 & 0.8\end{pmatrix}$, as shown in Fig. 4(c). By implementing the EM algorithm with the clustering number 4 and 100 different initials, there are 85 of 100 to obtain correct clustering results, as shown in Fig. 4(d). By implementing the FJ algorithm with the initial clustering number 20 and 100 different initializations, 92 of 100 obtain correct clustering results, as shown in Fig. 4(d). However, when the starting clustering number is changed to 30, only 81 of 100 obtain correct clustering results, as shown in Fig. 4(d). Then Figs. 4(e) and 4(f) are two other clustering results with $c^*=4$ obtained by the FJ algorithm. By implementing the robust EM to the data set as shown in Fig. 4(c) can obtain correct clustering results, as shown in Fig. 4(d) without an initial cluster number. If we generate 100 data sets with the same parameters and then implement the robust EM algorithm, 90 of 100 data sets obtain correct clustering results, as shown in Fig. 4(d).

Furthermore, we also compare our robust EM clustering algorithm with the FJ algorithm under the same initial values by starting cluster number $c_{initial}=20$ and 100 different random initial conditions. For the FJ algorithm, we know that there are 92 of 100 with the results of $c^*=4$ and the average iteration is 122.33. For our robust EM clustering algorithm, there are 94 of 100 with the results of $c^*=4$ and the average iteration is 39.95. We increase the cluster number $c_{initial}=30$ and also use 100 different random initial conditions. For the FJ algorithm, we find that there are only 81 of 100 with the results of $c^*=4$ and the average iteration is 181.41. For robust EM clustering algorithm, there are higher 96 of 100 with the results of $c^*=4$ and the average iteration is 39.34. Moreover, we generate 100 data sets and then use $c_{initial}=20$ and 100 different random initial conditions to each data set. We find that the FJ algorithm obtains the accuracy rate 0.7816 with the results of $c^*=4$. For our robust EM clustering algorithm, the accuracy rate 0.7954 with results of $c^*=4$. We then use $c_{initial}=30$ and 100 different random initial conditions to each

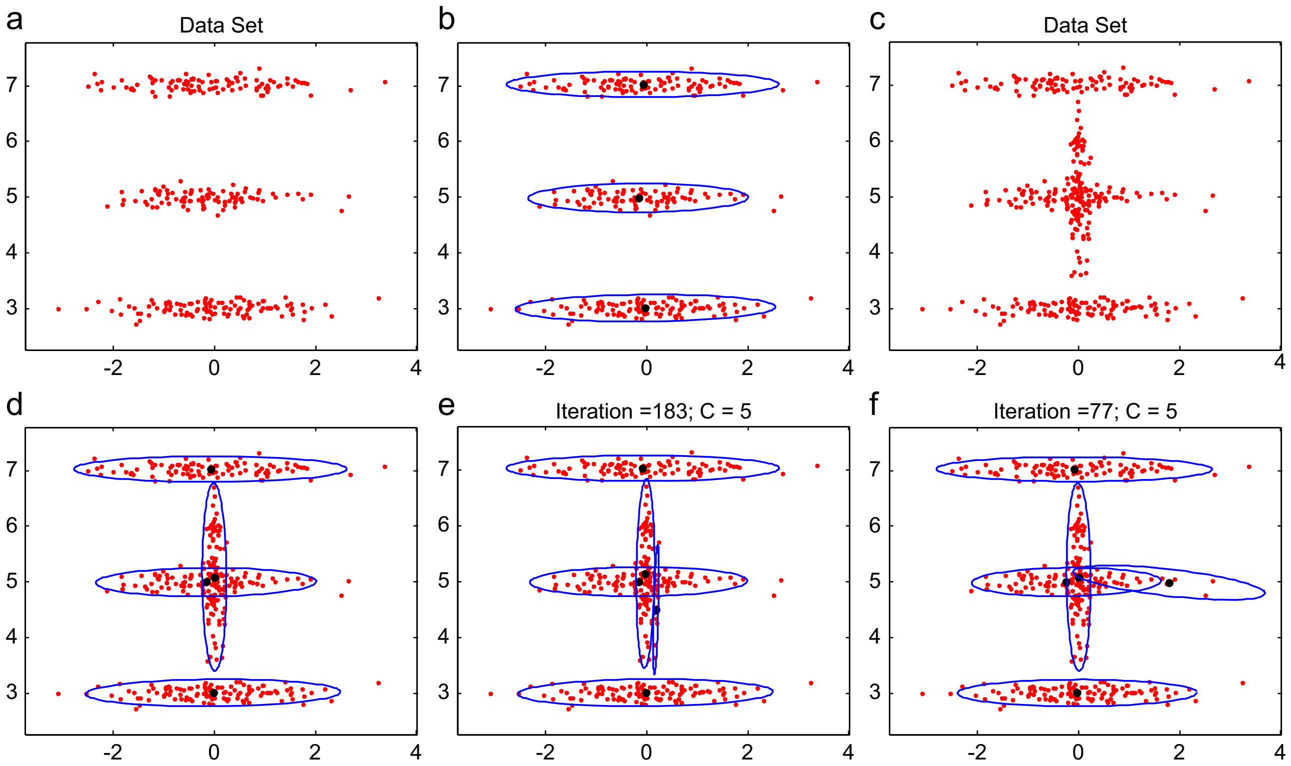

Fig. 4. (a) A well-separated data set; (b) clustering results of the EM algorithm; (c) a data set with adding one Gaussian data set to the data set as shown in (a); (d) clustering results of the EM algorithm when $c=3$ is given; (e) and (f) two other clustering results of Figueiredo and Jain's method with difference initializations.

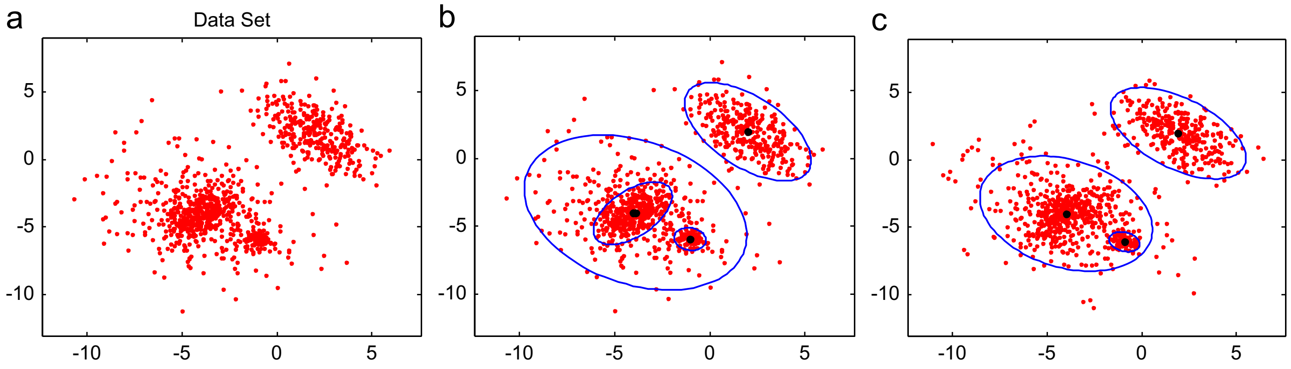

Fig. 5. (a) A two-dimensional, four-component Gaussian mixture data set with a sample size 1000; (b) correct clustering results for the data set (a) and (c) another clustering result with 3 clusters by the robust EM.

data set. We find that the FJ algorithm obtains the accuracy rate 0.5281 with the results of $c^{*}=4$. For our robust EM clustering algorithm, we have the accuracy rate 0.8592 with results of $c^{*}=4$.

Example 3. In this example, the data set is a two-dimensional, four-component Gaussian mixture distribution from Figueiredo and Jain [20] with parameters

$$\alpha_1=\alpha_2=\alpha_3=0.3,\quad \alpha_4=0.1;\quad \mu_1=\mu_2=(-4\quad -4)^T,$$

$$\mu_3=(2\quad 2)^T;\quad \mu_4=(-1\quad -6)^T,\quad \Sigma_1=\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix},$$

$$\Sigma_2=\begin{pmatrix}6 & -2\\ -2 & 6\end{pmatrix},\quad \Sigma_3=\begin{pmatrix}2 & -1\\ -1 & 2\end{pmatrix},\quad \Sigma_4=\begin{pmatrix}0.125 & 0\\ 0 & 0.125\end{pmatrix}.$$

We generate a data set from this Gaussian mixture model with a sample size $n=1000$, as shown in Fig. 5(a). This data set can be clustered into 4 clusters, as shown in Fig. 5(b). On the other hand, if we do not know the true clustering number, this data set can be viewed as two bigger clusters and one smaller cluster. Now we use Figueiredo and Jain's method for this data set with 100 different random initials. In our implementation, only 68 of 100 obtain the clustering results for 4 clusters, as shown in Fig. 5(b) and the other 32 of 100 obtain with more than 4 clusters. For this data set, implementing the proposed robust EM has the results shown in Fig. 5(b). If we further generate 100 data sets where each data set has 1000 sample points, we find that the proposed robust EM algorithm for these 100 data sets obtains 78 of 100 with 4 clusters, as shown in Fig. 5(b), 2 of 100 with 5 clusters, 16 of 100 with 3 clusters as shown in Fig. 5(c), and 4 of 100 with 2 clusters.

Furthermore, we also compare our robust EM clustering algorithm with the FJ algorithm with the same initial values by starting cluster number $c_{initial}=20$ and 100 different random initial conditions. For the FJ algorithm, we know that there are 68 of 100 with the results of $c^{*}=4$ and the average iteration is 198.62. For our robust EM clustering algorithm, there are 81 of

100 with the results of $c^*=4$ and the average iteration is 178.92. We also generate 100 data sets and then use $c_{initial}=20$ and 100 different random initial conditions to each data set. We find that the FJ algorithm obtains the accuracy rate 0.6113 with the results of $c^*=4$. For our robust EM clustering algorithm, we have the accuracy rate 0.6507 with results of $c^*=4$. As a whole, our robust EM clustering algorithm actually presents better than the FJ algorithm.

4. Examples and experimental results

In this section, some experimental examples are used to demonstrate the effectiveness of the proposed robust EM clustering algorithm. In all examples, we give $\varepsilon=0.0001$.

Example 4. The data set used in this example is from a one-dimensional, three-component Gaussian mixture distribution [22] with parameters

$$\alpha_1=\alpha_2=\alpha_3=1/3;\quad \mu_1=-11,\quad \mu_2=0,$$
$$\mu_3=13;\quad \sigma_1^2=4,\quad \sigma_2^2=16,\quad \sigma_3^2=9$$

We generate 1000 sample points, whose histogram is shown in Fig. 6(a). The robust EM algorithm is implemented for the data set. After 50 iterations, the cluster number $c^*=3$ is obtained and the estimates of parameters are as follows:

$$\hat{\alpha}_1=0.3475,\quad \hat{\alpha}_2=0.2989,\quad \hat{\alpha}_3=0.3536,\quad \hat{\mu}_1=-10.8118,$$
$$\hat{\mu}_2=0.0843,\quad \hat{\mu}_3=12.5974,\quad \hat{\sigma}_1^2=4.4722,\quad \hat{\sigma}_2^2=13.8800,$$
$$\hat{\sigma}_3^2=9.2170.$$

In Fig. 6(e), we use a red curve to show the real model and use the blue line to show the estimated model. We find that the estimated model is very close to the real model. In Fig. 6(f), the values of the objective function are demonstrated in that it has stably achieved its optimal value.

Example 5. In this example, we generate 1000 data points from a two-dimensional, five-component Gaussian mixture distribution with parameters

$$\alpha_1=\alpha_2=\alpha_3=\alpha_4=\alpha_5=0.2,\quad \mu_1=(0\quad 0)^T,\quad \mu_2=(0\quad 0)^T,$$
$$\mu_3=(-1.5\quad 1.5)^T$$
$$\mu_4=(1.5\quad 1.5)^T,\quad \mu_5=(0\quad -2)^T,\quad \Sigma_1=\begin{pmatrix}0.01 & 0\\ 0 & 1.25\end{pmatrix},$$
$$\Sigma_2=\begin{pmatrix}8 & 0\\ 0 & 8\end{pmatrix},\quad \Sigma_3=\begin{pmatrix}0.2 & 0\\ 0 & 0.015\end{pmatrix},\quad \Sigma_4=\begin{pmatrix}0.2 & 0\\ 0 & 0.015\end{pmatrix},$$
$$\Sigma_5=\begin{pmatrix}1 & 0\\ 0 & 0.2\end{pmatrix}$$

where Geva [23] used it to imitate the so-called “face” data as shown in Fig. 7(a). We implement the robust EM algorithm for the data set. The clustering results after different iterations are shown in Figs. 7(b)–7(e), respectively, indicating that the number of clusters will gradually decrease. The algorithm is convergent after 149 iterations with the optimal cluster number $c^*=5$, as shown in Fig. 7(e). In Fig. 7(f), we can see that the curve for the values of the objective function becomes horizontal, which is because the number of clusters is stable and β is equal to zero. The estimates of parameters are finally as follows:

$$\hat{\alpha}_1=0.2148,\quad \hat{\alpha}_2=0.1731,\quad \hat{\alpha}_3=0.2024,$$
$$\hat{\alpha}_4=0.1921,\quad \hat{\alpha}_5=0.2176$$
$$\hat{\mu}_1=\begin{pmatrix}0.0010\\ -0.0348\end{pmatrix},\quad \hat{\mu}_2=\begin{pmatrix}-0.0549\\ -0.1078\end{pmatrix},\quad \hat{\mu}_3=\begin{pmatrix}-1.4633\\ 1.5139\end{pmatrix},$$
$$\hat{\mu}_4=\begin{pmatrix}1.5396\\ 1.4893\end{pmatrix},\quad \hat{\mu}_5=\begin{pmatrix}-0.0799\\ -1.9000\end{pmatrix},\quad \hat{\Sigma}_1=\begin{pmatrix}0.0083 & 0.0008\\ 0.0008 & 1.3381\end{pmatrix}$$
$$\hat{\Sigma}_2=\begin{pmatrix}6.5406 & 0.0327\\ 0.0327 & 8.4815\end{pmatrix},\quad \hat{\Sigma}_3=\begin{pmatrix}0.1996 & -0.0084\\ -0.0084 & 0.0145\end{pmatrix},$$

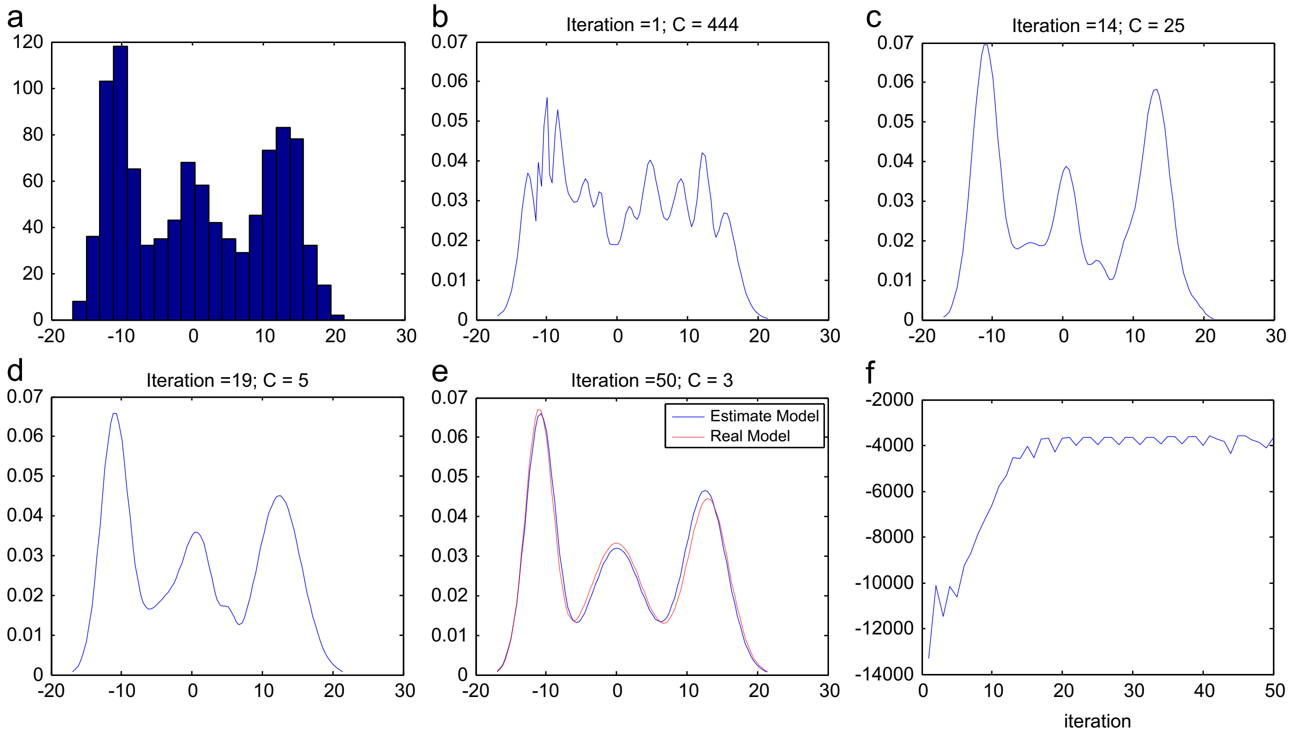

Fig. 6. (a) Histogram of the 1000 data points; (b) estimating model after 1 iteration with $c=444$; (c) estimating model after 14 iteration with $c=25$; (d) estimating model after 19 iteration with $c=5$; (e) estimating model after 50 iteration in convergence with $c^*=3$ in which the blue curve is our estimating model and the red curve is the real model. (f) The values of the objective function.

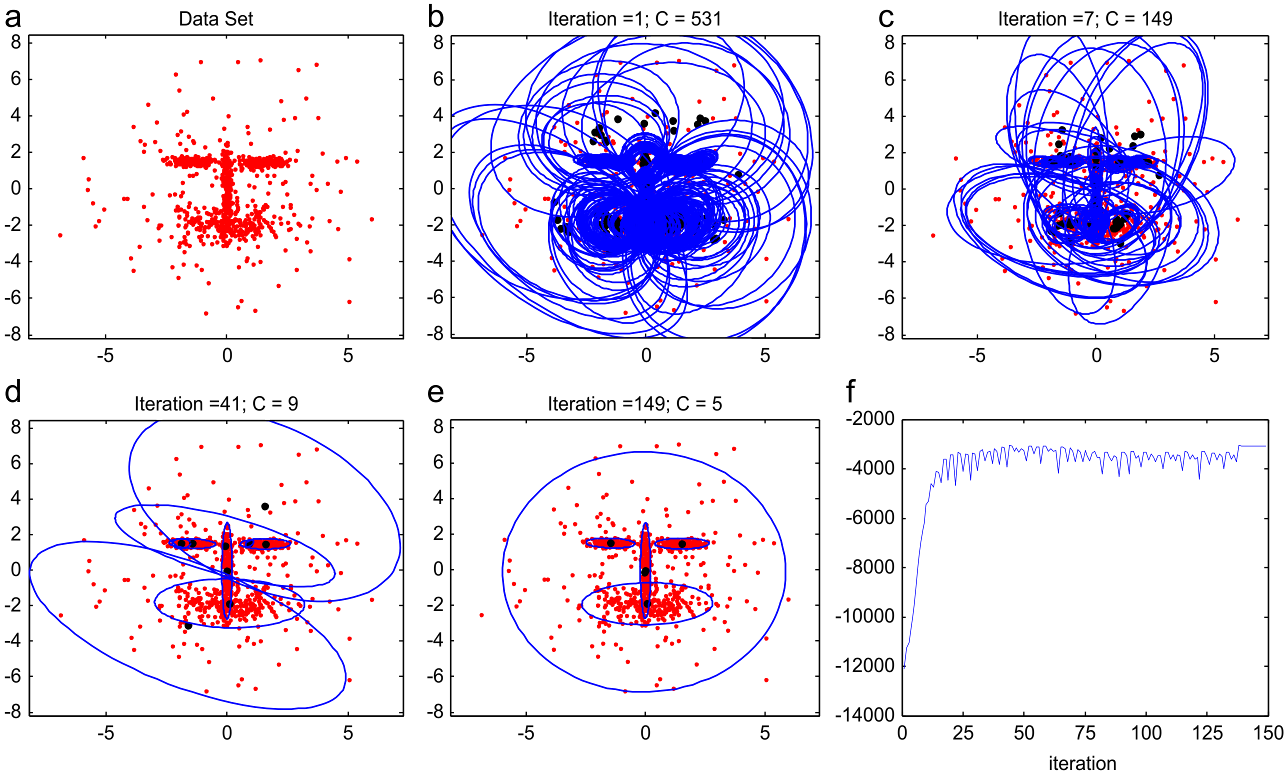

Fig. 7. (a) The original data set; (b) results with $c=531$ after one iteration; (c) results with $c=149$ after 7 iterations; (d) results with $c=41$ after 9 iterations; (e) results with $c^{*}=5$ after 149 iterations to convergence and (f) the values of objective function.

$$\hat{\Sigma}_4 = \begin{pmatrix} 0.2438 & 0.0013 \\ 0.0013 & 0.0167 \end{pmatrix}, \quad \hat{\Sigma}_5 = \begin{pmatrix} 1.3988 & 0.0238 \\ 0.0238 & 0.2591 \end{pmatrix}.$$

If we further use the same parameters to generate 100 data sets, we find that the robust EM algorithm obtain 90 of 100 data sets to have good clustering results, as shown in Fig. 7(e) with $c^{*}=5$.

Example 6. In this example, we use an artificial data set as shown in Fig. 8(a). There are 16 clusters and each cluster has 50 data points. Even though these 16 clusters seem to be clearly separated, if we use the EM algorithm with the cluster number $c=16$, it always obtains a bad clustering result shown in Fig. 8(b) unless good initial values are given for the EM algorithm. When we implement the robust EM algorithm for the data set, we find that not only can we obtain correct clustering number with $c^{*}=16$, but we also can obtain good clustering results after 31 iterations, as shown in Fig. 8(f).

Example 7. In this example, we use the robust EM algorithm for the data set shown in Fig. 9(a), which has three dimensions and nine clusters with different shapes. After implementing the robust EM algorithm with 140 iterations, we obtain final clustering results in convergence, as shown in Fig. 9(b). The clustering results by the robust EM algorithm are good even though the data set has different cluster shapes.

Example 8. The real data set of flea beetle data from Lubischew [24] is used in this example. The data set has 74 data points with three species: *concinna* (21), *heikertingeri* (31) and *heptapotamica* (22). Each data point was obtained by measuring two characteristics of a beetle: the maximal width of the aedeagus in the fore-part in microns and the front angle of the aedeagus (1 unit=7.5°). The flea beetle data set is shown in Fig. 10(a). By implementing the EM algorithm for this data set with 100 random initials, we have 49 of 100 with the clustering results shown in Fig. 10(b), 20 of 100 with the clustering results shown in Fig. 10(c), and 31 of 100 with various other clustering results. By implementing the robust EM algorithm, we obtain the clustering results shown in Fig. 10(d) with $c^{*}=3$, where there is only 1 *concinna* incorrectly classified as *heikertingeri*, as shown in Fig. 10(e).

Example 9. In this example, the real data set with three-dimensional data from Reaven and Miller [25] is considered. There are 145 observations of diabetes patients and three measurements on each patient: the plasma glucose response to oral glucose, the plasma insulin response to oral glucose, and the degree of insulin resistance. They were clinically classified into three groups: *normal* (76), *chemical* diabetes (36), and *overt* diabetes (33). The classified 3D graph is shown in Fig. 11(a). By implementing the robust EM algorithm for this data set, we get the clustering results shown in Fig. 11(b) in convergence after 133 iterations with $c^{*}=3$. There are 20 misclassifications out of 145 observations.

Finally, we analyze the computational complexity for the proposed robust EM algorithm. In fact, the robust EM has the same computational complexity as the original EM. The difference of computational complexity between these two methods is the initial cluster number c, where c varies from n to c_{final}. The computational complexity is calculated per iteration from two parts, E-step and M-step. In the E-step, the computational complexity for computing the $\hat{z}_{ki}$ is $O(nc)$. In the M-step, the computational complexities for computing the proportions α_k and the means μ_k are both $O(c)$. For computing the covariance matrices Σ_k, the computational complexity is $O(nd^2c)$. Thus, the computational complexity in each iteration is $O(nc+c+nd^2c)=O(nc(1+d^2)+c)$ where c varies from n to c_{final}. Although the robust EM uses the number of data points $c=n$ as the cluster number in the beginning of implementation, the time per iteration will decrease rapidly after several iterations. This is because the cluster number c decreases rapidly by discarding

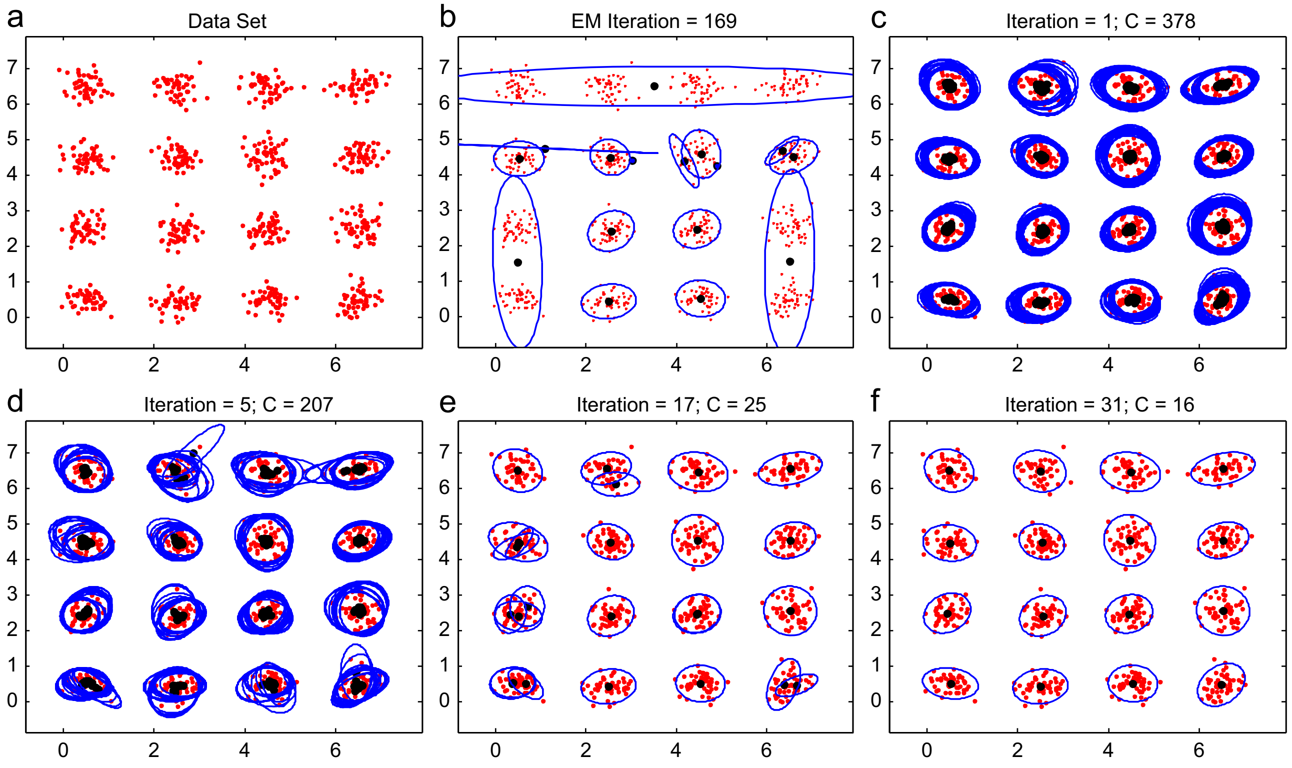

Fig. 8. (a) The original data set; (b) clustering results of the EM algorithm with random initializations; (c)–(e) clustering results of the robust EM after 1,5, and 17 iterations and (f) clustering results in convergence after 31 iterations.

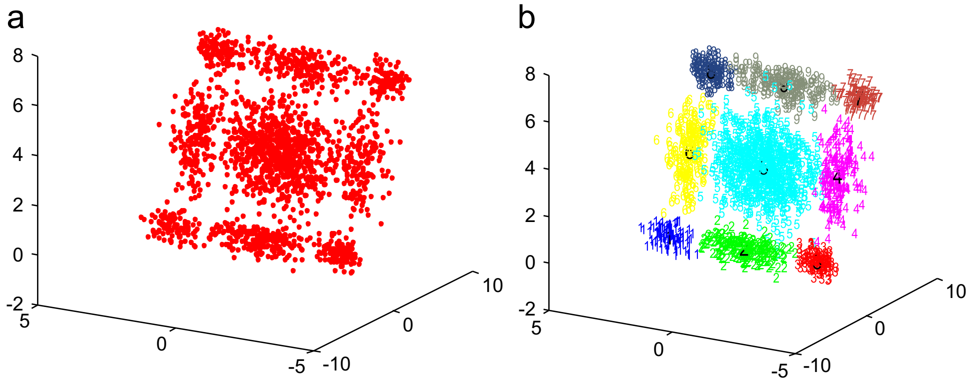

Fig. 9. (a) The three-dimensional data set with nine clusters and (b) clustering results by the robust EM algorithm in convergence after 140 iterations with $c^*=9$.

those clusters with $\alpha_k^{(t)} \leq 1/n$ during implementation. To demonstrate this phenomenon, we show the computation time in seconds per iteration for the data sets of Example 4, 5 and 7 as shown in Fig. 12 and Table 1. We find that the computation time decreases rapidly after the 10th iteration for Examples 4, 5 and 7.

5. Conclusions and discussion

We know that the EM algorithm is sensitive to initial values. In this paper we propose a new schema for the EM without initialization. The proposed robust EM algorithm for Gaussian mixture models uses all data points as initials to solve the problem of choosing initial values. We then use a penalty term to construct a competition schema. When a cluster during implementing the algorithm has an illegitimate proportion, we can discard it based on the construction. If the cluster number does not decrease, we achieve an appropriate cluster number and also parameter estimations so that an optimal cluster number can be automatically found according to the structure of data. Several examples with numerical and real data sets demonstrate the superiority of the proposed robust EM clustering algorithm for Gaussian mixture models. We had mentioned that some convergence properties of the EM algorithm (for Gaussian mixtures) had been constructed by Wu [17] and Xu and Jordan [18]. Although we had considered the maximum solution of Eq. (12) with the penalized function $J(\alpha,\theta)$ based on the necessary conditions of the Lagrangian $\tilde{J}$, we had added some formula and also adjusted $\alpha_{k'}$ and $\hat{z}_{k'i}$ to have a competition schema setting such that the algorithm can automatically reduce the number of clusters and also simultaneously get the estimates of parameters. In this case, the convergence properties of our robust EM clustering algorithm cannot be proved by a similar way as Wu [17] and Xu and Jordan [18]. However, we may borrow some results from dynamical systems. If we can reform the

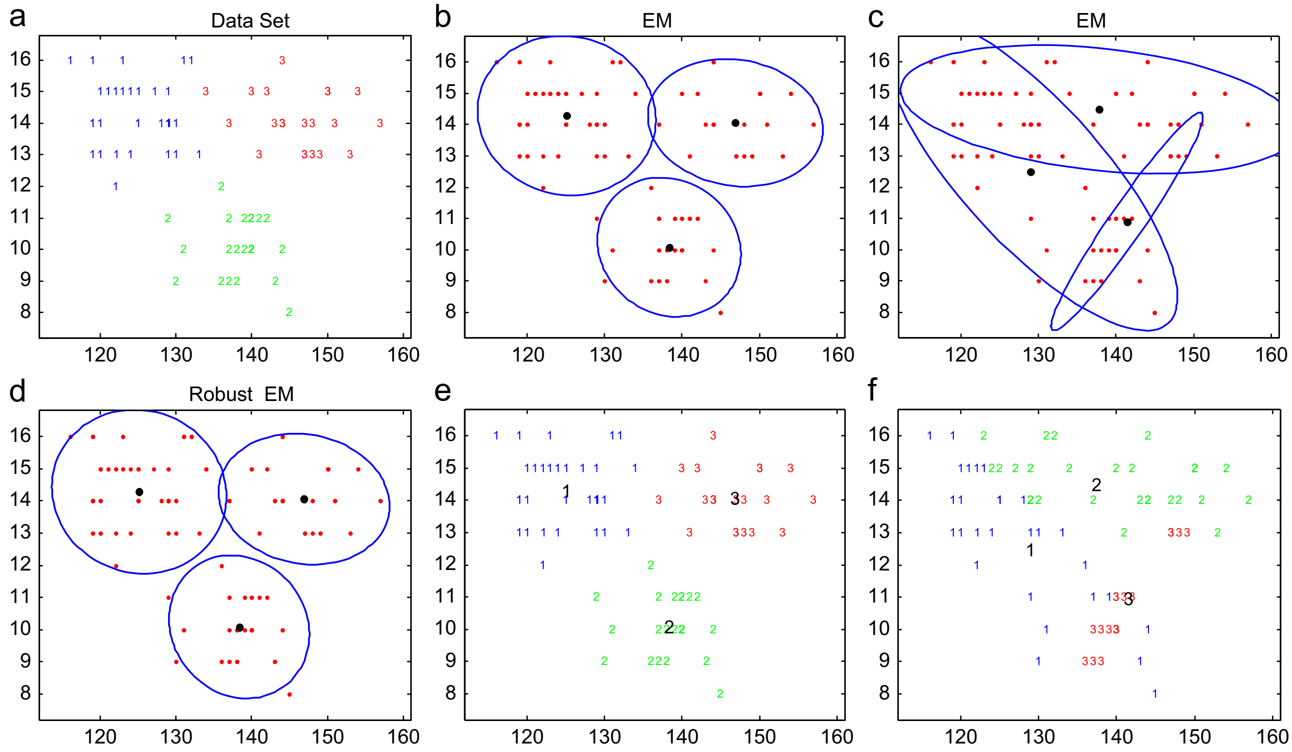

Fig. 10. (a) The original flea beetle data set where "1" denotes the cluster of heikertingeri, "2" denotes the cluster of heptapotamica and "3" denotes the cluster of concinna; (b) and (c) are two different clustering results by the EM algorithm with random initials; (d) clustering results of the robust EM; (e) the identified clusters of (b) and (d). (f) The identified clusters of (c).

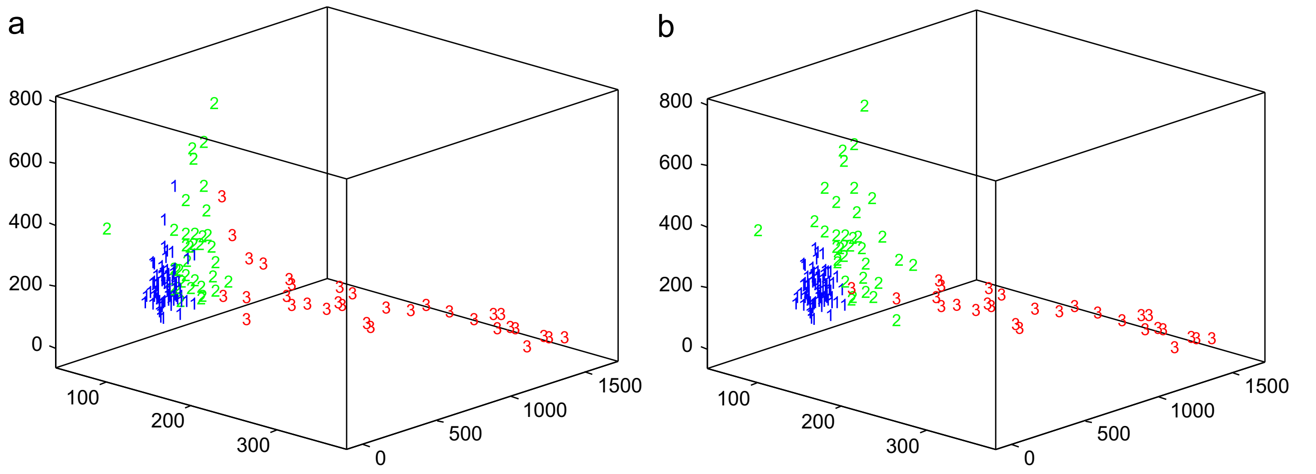

Fig. 11. (a) The 3D plot of the diabetes data set where '1' denotes the 'normal' class, '2' denotes the 'chemical' class, and '3' denotes the 'overt' class; (b) clustering results of the robust EM algorithm.

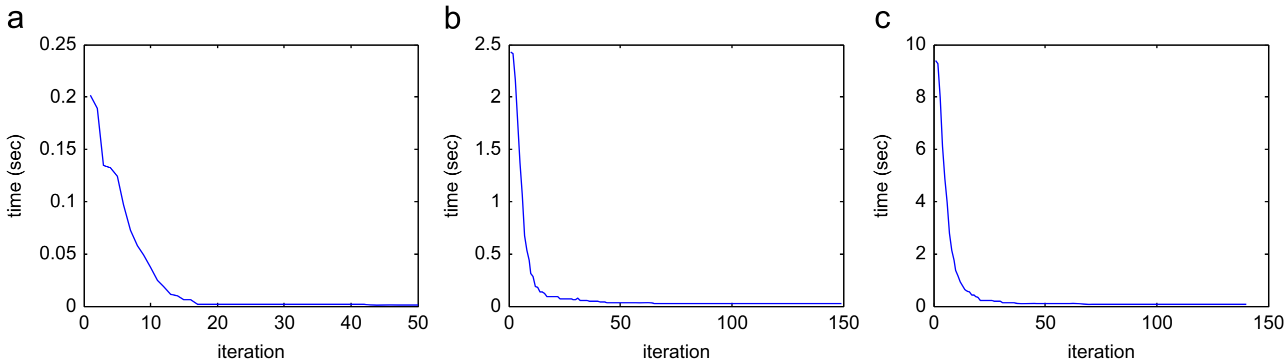

Fig. 12. Plots of per iteration time as implementing the robust EM for the data sets in (a) Example 4; (b) Example 5 and (c) Example 7.

Table 1
Running time (s) of the first ten iterations for the data sets in Examples 4, 5 and 7 with different dimensions.

Data set	Iteration									
	1	2	3	4	5	6	7	8	9	10
Example 4 ($d=1$)	0.202	0.189	0.135	0.133	0.124	0.097	0.073	0.058	0.049	0.037
Example 5 ($d=2$)	2.428	2.412	2.164	1.776	1.373	1.065	0.673	0.533	0.438	0.313
Example 7 ($d=3$)	9.385	9.272	7.976	6.119	4.895	3.956	2.784	2.149	1.743	1.373

proposed robust EM algorithm in a language of dynamical systems, we may obtain some convergence properties of the proposed robust EM algorithm. This will be our future research topic.

On the other hand, Gaussian distribution is not robust for outliers. Some distributions, such as t-distribution and Pearson type VII distribution, are more robust to outliers than Gaussian distribution. In the literature, there are several ways to reconstruct the EM as a robust clustering algorithm for outliers. There are some by replacing a mixture of Gaussian distributions with a mixture of *t*-distributions [26,27], with a mixture of skew-normal distributions [28] or with a mixture of Pearson type VII distributions [29]. Some others consider by using robust estimators, such as the trimmed likelihood estimator [30], trimmed *k*-means with trimmed data set [31] or modified *t*-factor analyzer [32], and so forth. We know that the proposed robust EM algorithm is robust for initial values and cluster number. In our further work, we will advance our robust EM algorithm as a more robust algorithm for outliers. We may consider application of our robust EM algorithm to the mixture of *t*-distributions and Pearson type VII distributions, or further consider those robust estimators.

Acknowledgments

The authors would like to thank the anonymous referees for their helpful comments in improving the presentation of this paper. This work was supported in part by the National Science Council of Taiwan, under Grant NSC-99-2118-M-033-004-MY2.

Appendix

To derive the update equation for α_k under the constraint $\sum_{k=1}^{c}\alpha_k=1$, the Lagrangian $\tilde{J}$ of Jshould be

$$\tilde{J}(\alpha,\theta)=\sum_{i=1}^{n}\sum_{k=1}^{c}\hat{z}_{ki}\ln[\alpha_k f(x_i;\theta_k)]+\beta\sum_{i=1}^{n}\sum_{k=1}^{c}\alpha_k\ln\alpha_k-\lambda\left(\sum_{k=1}^{c}\alpha_k-1\right)$$

We take the first derivative of the Lagrangian $\tilde{J}$ with respect to α_k and set it to be zero. We derive an equation for α_k, as follows:

$$\frac{\sum_{i=1}^{n}\hat{z}_{ki}}{\alpha_k}+n\beta\ln\alpha_k+n\beta-\lambda=0$$

Then we have

$$\sum_{i=1}^{n}\hat{z}_{ki}+n\beta\alpha_k\ln\alpha_k+n\beta\alpha_k-\lambda\alpha_k=0 \tag{A1}$$

$$\sum_{k=1}^{c}\sum_{i=1}^{n}\hat{z}_{ki}+n\beta\sum_{k=1}^{c}\alpha_k\ln\alpha_k+n\beta\sum_{k=1}^{c}\alpha_k-\lambda\sum_{k=1}^{c}\alpha_k=0$$

$$\lambda=n+n\beta\sum_{k=1}^{c}\alpha_k\ln\alpha_k+n\beta \tag{A2}$$

Embedding (A2) into Eq. (A1), we have

$$\sum_{i=1}^{n}\hat{z}_{ki}+n\beta\alpha_k\ln\alpha_k+n\beta\alpha_k-(n+n\beta\sum_{s=1}^{c}\alpha_s\ln\alpha_s+n\beta)\alpha_k=0$$

$$\sum_{i=1}^{n}\hat{z}_{ki}+n\beta\alpha_k\ln\alpha_k-n\alpha_k-n\beta\alpha_k\sum_{s=1}^{c}\alpha_s\ln\alpha_s=0$$

Then the updated equation for α_k is

$$\alpha_k^{(new)}=\frac{\sum_{i=1}^{n}\hat{z}_{ki}}{n}+\beta\alpha_k^{(old)}(\ln\alpha_k^{(old)}-\sum_{s=1}^{c}\alpha_s^{(old)}\ln\alpha_s^{(old)})$$

References

[1] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, NJ, 1988.

[2] G.J. McLachlan, K.E. Basford, Mixture Models: Inference and Applications to clustering, Marcel Dekker, New York, 1988.

[3] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 1990.

[4] R. Duda, P. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[5] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), Journal of the Royal Statistical Society-Series B 39 (1977) 1–38.

[6] J. MacQueen, Some methods for classification and analysis of multivariate observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281-297, University of California Press, 1967.

[7] D. Pollard, Quantization and the method of *k*-means, IEEE Transactions on Information Theory 28 (1982) 199–205.

[8] J.A. Guesta-Albertos, A. Gordaliza, C. Matran, Trimmed *k*-means: an attempt to robustify quantizers, Annals of Statistics 25 (1997) 553–576.

[9] L.A. Garcia-Escudero, A. Gordaliza, Robustness properties of *k*-means and trimmed k-means, Journal of the American Statistical Association 94 (1999) 956–969.

[10] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[11] K.L. Wu, M.S. Yang, Alternative *c*-means clustering algorithms, Pattern Recognition 35 (2002) 2267–2278.

[12] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (1995) 790–799.

[13] K.L. Wu, M.S. Yang, Mean shift-based clustering, Pattern Recognition 40 (2007) 3035–3052.

[14] C. Biernacki, G. Celeux, G. Govaert, Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models, Computational Statistic & Data Analysis 41 (2003) 561–575.

[15] C.K. Reddy, H.D. Chiang, B. Rajaratnam, TRUST-TECH-based expectation maximization for learning finite mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 1146–1157.

[16] S. Richardson, P.J. Green, On Bayesian analysis of mixtures with an unknown number of components, Journal of the Royal Statistical Society-Series B 59 (1997) 731–758.

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

[18] L. Xu, M.I. Jordan, On convergence properties of the EM algorithm for Gaussian mixtures, Neural Computation 8 (1996) 129–151.

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

[20] M.A.T. Figueiredo, A.K. Jain, Unsupervised learning of finite Mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 381–396.

[21] G. Celeux, S. Chrétien, F. Forbes, A. Mkhadri, A component-wise EM algorithm for mixtures, Journal of Computational and Graphical Statistics 10 (2001) 697–712.

[22] N. Ueda, R. Nakano, Deterministic annealing EM algorithm, Neural Networks 11 (1998) 271–282.

[23] A.B. Geva, Hierarchical unsupervised fuzzy clustering, IEEE Transactions on Fuzzy Systems 7 (1999) 723–733.
[24] A.A. Lubischew, On the use of discriminant functions in taxonomy, Biometrics 18 (1962) 455–477.
[25] G.M. Reaven, R.G. Miller, An attempt to define the nature of chemical diabetes using a multidimensional analysis, Diabetologia 16 (1979) 17–24.
[26] D. Peel, G.J. MacLaren, Robust mixture modeling using the t-distribution, Statistics and Computing 10 (2000) 339–348.
[27] K. Lo, R. Gottardo, Flexible mixture modeling via the multivariate t distribution with the Box-Cox transformation: an alternative to the skew-*t* distribution, Statistics and Computing 22 (2012) 33–52.
[28] R.M. Basso, V.H. Lachos, C.R.B. Cabral, P. Ghosh, Robust mixture modeling based on scale mixtures of skew-normal distributions, Computational Statistics and Data Analysis 54 (2010) 2926–2941.
[29] J. Sun, A. Kaban, J.M. Garibaldi, Robust mixture clustering using Pearson type VII distribution, Pattern Recognition Letters 31 (2010) 2447–2454.
[30] N. Neykova, P. Filzmoserb, R. Dimovac, P. Neytcheva, Robust fitting of mixtures using the trimmed likelihood estimator, Computational Statistics & Data Analysis 52 (2007) 299–308.
[31] J.A. Cuesta-Albertos, C. Matrán, A. Mayo-Iscar, Robust estimation in the normal mixture model based on robust clustering, Journal of the Royal Statistical Society B 70 (2008) 779–802.
[32] J.L. Andrews, P.D. McNicholas, Mixtures of modified *t*-factor analyzers for model-based clustering, classification, and discriminant analysis, Journal of Statistical Planning and Inference 141 (2011) 1479–1486.

Miin-Shen Yang received the BS degree in mathematics from the Chung Yuan Christian University, Chung-Li, Taiwan, in 1977, the MS degree in applied mathematics from the National Chiao-Tung University, Hsinchu, Taiwan, in 1980, and the PhD degree in statistics from the University of South Carolina, Columbia, USA, in 1989.

In 1989, he joined the faculty of the Department of Mathematics in the Chung Yuan Christian University as an Associate Professor, where, since 1994 he has been a Professor. From 1997 to 1998, he was a Visiting Professor with the Department of Industrial Engineering, University of Washington, Seattle. During 2001–2005, he was the Chairman of the Department of Applied Mathematics in the Chung Yuan Christian University. His current research interests include applications of statistics, fuzzy clustering, pattern recognition, machine learning, and neural fuzzy systems.

Dr. Yang is an Associate Editor of the IEEE Transactions on Fuzzy Systems, and an Associate Editor of the Applied Computational Intelligence and Soft Computing.

Chien-Yo Lai received the BS degree in applied mathematics from the Feng Chia University, Taichung, Taiwan, in 2000, the MS and PhD degrees in applied mathematics from the Chung Yuan Christian University, Chung-Li, Taiwan, in 2002 and 2010. He is currently a part-time Assistant Professor at the Department of Applied Mathematics in the Chung Yuan Christian University, Taiwan. His research interests include cluster analysis and pattern recognition.

Chih-Ying Lin received the BS degree in applied mathematics from the Chung Yuan Christian University, Chung-Li, Taiwan, in 2003, the MS degrees in applied mathematics from the National Dong Hwa University, Hualien, Taiwan, in 2005. He is a Ph.D. student in the Department of Applied Mathematics at Chung Yuan Christian University, Chung-Li, Taiwan. His research interests include cluster analysis, and fuzzy data analysis.