Robust Estimation in Gaussian Mixtures Using Multiresolution $\boldsymbol{K}$d-trees

Shu-Kay Ng and Geoffrey J. McLachlan

Department of Mathematics, University of Queensland,
Brisbane, QLD 4072, Australia
skn@maths.uq.edu.au and gjm@maths.uq.edu.au

Abstract. For many applied problems in the context of clustering via mixture models, the estimates of the component means and covariance matrices can be affected by observations that are atypical of the components in the mixture model being fitted. In this paper, we consider for Gaussian mixtures a robust estimation procedure using multiresolution kd-trees. The method provides a fast EM-based approach to the fitting of Gaussian mixtures in applications to huge data sets. In addition, a robust estimation against outliers in fitting Gaussian mixtures is achieved by giving reduced weight to observations that are atypical of a component. The method is illustrated using real and simulated data.

1 Introduction

With a Gaussian mixture model approach to clustering, it is assumed that the observed p-dimensional vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are from a mixture of, say g, components in some unknown proportions $\pi_1, \ldots, \pi_g$ that sum to one. That is, each data point is taken to be a realization of the mixture probability density function,

$$f(\boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

where $\phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the p-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Here the vector $\boldsymbol{\Psi}$ of unknown parameters consists of $\pi_1, \ldots, \pi_{g-1}$, the elements of $\boldsymbol{\mu}_i$, and the distinct elements of $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. The vector $\boldsymbol{\Psi}$ can be estimated by the maximum likelihood method via the expectation-maximization (EM) algorithm [1].

Within the EM framework, each $\boldsymbol{x}_j$ is conceptualized to have arisen from one of the g components. We let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ denote the unobservable component-indicator vectors, where the ith element z_{ij} of $\boldsymbol{z}_j$ is taken to be one or zero according as the jth data point $\boldsymbol{x}_j$ does or does not come from the ith component. We put $\boldsymbol{z} = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)^T$ where the superscript T denotes vector transpose. An outright clustering of the data into g clusters can be obtained by assigning the jth data point to the component to which it has the highest estimated posterior probability of belonging [2, Section 1.15].

For many applied problems in the context of clustering via mixture models, the estimates of the component means and covariance matrices can be affected by observations that are atypical of the components in the mixture model being fitted. In this paper, we consider the use of multiresolution kd-trees (mrkd-trees) to provide a robust approach to the fitting of Gaussian mixtures. With the mrkd-tree approach, "close-by" observations are grouped into tree-nodes. The method thus speeds up the implementation of the EM algorithm in the fitting of Gaussian mixtures to huge low-dimensional data sets [3, 4]. Moreover, observations that are atypical of a component are being given reduced weight in the calculation of its parameters. The method thus also provides a robust estimation against outliers in fitting Gaussian mixtures. Although the mrkd-tree approach has been considered as an approximate method (see Section 2), Ng and McLachlan [4] has shown that mrkd-tree-based algorithms can converge to essentially the same maximum log likelihood value as the EM algorithm.

Gaussian mixtures are increasingly being adopted in applications in image processing contexts, such as the segmentation of magnetic resonance (MR) images [5, 6]. A typical three-dimensional (3D) multispectral MR image consists of low-dimensional ($p \leq 3$) feature vectors of intensities measured on over ten millions voxels. Thus the aim of the proposed method is to provide a fast EM-based mixture model approach to segment MR images and also a robust estimation of image parameters against the intensity inhomogeneity due to acquisition equipments.

The paper is organized as follows: Section 2 introduces a sparse, incremental (SPIEM) mrkd-tree algorithm proposed by [4] to speed up the EM algorithm. In Section 3, we consider a robust procedure which identifies tree-nodes as three different types and gives different weights on them in the calculation of parameters. In Section 4, the performance of the proposed method is illustrated using some simulated and real data. Section 5 ends the paper with some discussion.

2 A SPIEM mrkd-tree algorithm

The use of a mrkd-tree has been proposed by Moore [3] to speed up the EM algorithm. Here kd stands for k-dimensional where, in our notation, $k = p$, the dimension of a feature vector $\boldsymbol{x}_j$. The kd-tree is a binary tree that recursively splits the whole set of data into regions. Each node in the kd-tree includes a bounding box that specifies a subset of the data points and the root node owns all the data. The children of a node are smaller bounding boxes, generated by splitting along the parent node's widest dimension. The mrkd-tree is constructed top-down, starting from the root node and the splitting procedure continues until the range of data points in the widest dimension of a descendant node is smaller than some threshold γ. This node is then declared to be a leaf-node and is left unsplit. In this paper, we take γ to be 0.3% of the range in the splitting dimension of the whole data set [4].

For Gaussian mixtures, it is computationally advantageous to work in terms of the sufficient statistics [7]. With the help of the multiresolution data structure built up by the kd-tree, the computation of the current conditional expectations of the sufficient statistics in the E-step can be restructured as follows on the $(k + 1)$th scan of the EM algorithm. Let n_L be the total number of leaf nodes. For the mth leaf node LN_m $(m = 1, \ldots, n_L)$, the conditional expectations of the sufficient statistics are simplified by treating all the data points in it to have the same posterior probabilities $\tau_i(\bar{\boldsymbol{x}}_m;\boldsymbol{\Psi}^{(k)})$ calculated at the mean, where

$$\tau_i(\bar{\boldsymbol{x}}_m;\boldsymbol{\Psi}^{(k)}) = \pi_i^{(k)}\phi(\bar{\boldsymbol{x}}_m;\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Sigma}_i^{(k)})/\sum_{l=1}^{g}\pi_l^{(k)}\phi(\bar{\boldsymbol{x}}_m;\boldsymbol{\mu}_l^{(k)},\boldsymbol{\Sigma}_l^{(k)}) \quad (i=1,\dots,g)$$

and $\bar{\boldsymbol{x}}_m$ is the mean of data points belonging to the mth node. The contribution of the mth leaf node LN_m $(m=1,\dots,n_L)$ to the conditional expectations of the sufficient statistics is given as, for $i=1,\dots,g$,

$$\begin{aligned}
T_{i1,m}^{(k)} &= \tau_i(\bar{\boldsymbol{x}}_m;\boldsymbol{\Psi}^{(k)})n_m\\
\boldsymbol{T}_{i2,m}^{(k)} &= \tau_i(\bar{\boldsymbol{x}}_m;\boldsymbol{\Psi}^{(k)})n_m\bar{\boldsymbol{x}}_m\\
\boldsymbol{T}_{i3,m}^{(k)} &= \tau_i(\bar{\boldsymbol{x}}_m;\boldsymbol{\Psi}^{(k)})\sum_{j\in LN_m}\boldsymbol{x}_j\boldsymbol{x}_j^T,
\end{aligned}$$

where n_m is the number of data points in the mth node $(m=1,\dots,n_L)$. The conditional expectations of the sufficient statistics are approximated as

$$T_{i1}^{(k)} = \sum_{j=1}^{n}\tau_{ij}^{(k)} \approx \sum_{m=1}^{n_L}T_{i1,m}^{(k)} \tag{2}$$

$$\boldsymbol{T}_{i2}^{(k)} = \sum_{j=1}^{n}\tau_{ij}^{(k)}\boldsymbol{x}_j \approx \sum_{m=1}^{n_L}T_{i2,m}^{(k)} \tag{3}$$

$$\boldsymbol{T}_{i3}^{(k)} = \sum_{j=1}^{n}\tau_{ij}^{(k)}\boldsymbol{x}_j\boldsymbol{x}_j^T \approx \sum_{m=1}^{n_L}T_{i3,m}^{(k)}, \tag{4}$$

where $\tau_{ij}^{(k)} = E(Z_{ij} \mid \boldsymbol{x};\boldsymbol{\Psi}^{(k)})$ is the current estimate of the posterior probability that $\boldsymbol{x}_j$ comes from the ith component $(i=1,\dots,g;\ j=1,\dots,n)$.

The M-step updates the estimates as follows:

$$\pi_i^{(k+1)} = T_{i1}^{(k)}/n, \tag{5}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \boldsymbol{T}_{i2}^{(k)}/T_{i1}^{(k)}, \tag{6}$$

$$\boldsymbol{\Sigma}_i^{(k+1)} = \left\{\boldsymbol{T}_{i3}^{(k)} - T_{i1}^{(k)^{-1}}\boldsymbol{T}_{i2}^{(k)}\boldsymbol{T}_{i2}^{(k)^T}\right\}/T_{i1}^{(k)}. \tag{7}$$

It is noted that the calculation of the sufficient statistics in (2) to (4) is approximate. In practice, the leaf nodes should be very small (or γ small) in order that the simplified equations (2) to (4) be applicable. However, in this situation, n_L will be close to the number of data points n, and hence there is very little computational gain over the standard EM algorithm.

Thus, a further (pruning) step is proposed to reduce the computational time [3]. For each component i at a given node $(i=1,\dots,g)$, the minimum and maximum values that any data point in the node can have for its current posterior probabilities are computed. Denote these limiting values $\tau_{i,min}$ and $\tau_{i,max}$,

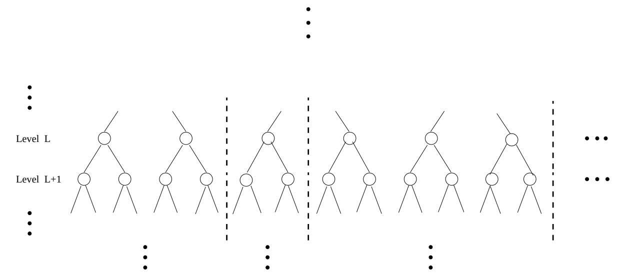

Fig. 1. The SPIEM–*k*d-tree algorithm. Partition of nodes at level L into blocks

respectively. If the differences between them for all $i = 1, \ldots, g$ are small and satisfy a pruning criterion (see below), then the node is treated as if it is a (pseudo) leaf node. Hence its descendants need not be searched at this scan and time is saved. Let $\tau_{i,total}$ be the sum of posterior probabilities of the ith component membership for all the data points. We prune the mth node if

1. $n_m(\tau_{i,max} - \tau_{i,min}) < 0.01\tau_{i,total} \quad \forall i = 1, \ldots, g$, and
2. $\log(\sum_{i=1}^{g} \pi_i\phi_{i,max} / \sum_{i=1}^{g} \pi_i\phi_{i,min}) < 0.5 \mid \log \sum_{i=1}^{g} \pi_i\phi(\bar{\boldsymbol{x}}_m; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid$,

where $\phi_{i,max}$ and $\phi_{i,min}$ are the upper and lower bound on the ith component-conditional density, respectively. For $p \leq 3$, the limiting values of τ_i are obtained using an analytical geometry approach. The idea is to transform the data points by a matrix of normalized eigenvectors so that the covariance matrix becomes an identity matrix and hence the Mahalanobis squared distance becomes the Euclidean squared distance; see [4] for details.

In [4], a SPIEM *mrk*d-tree algorithm is proposed to speed up the EM algorithm for fitting Gaussian mixtures. With this algorithm, the nodes at a predetermined level, say L, of the *k*d-tree are divided into B blocks and a "partial" E-step is implemented by searching down from only a block of nodes at level L at a time before the next M-step is performed (Fig. 1). Here the number of blocks B is chosen based on the simple rule proposed in [7]. The argument for improved convergence rate is that the algorithm exploits new information more quickly rather than waiting for a complete scan of all nodes before parameters are updated by an M-step. Moreover, component-posterior probabilities that are below a specified threshold are held fixed while those for the remaining components in the mixture are updated. Thus, instead of considering all g components, it is possible to "freeze" those $\tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k)})$ that are close to zero and save time.

To examine the SPIEM *mrk*d-tree algorithm more closely, let A_m ($m = 1, \ldots, n_{PL}$) be a subset of $\{1, \ldots, g\}$ which component-posterior probability of the mth pseudo-leaf node is close to zero, say less than 0.005, and hence is held fixed [4]. Here n_{PL} is the number of pseudo-leaf nodes at the current scan. Let $\boldsymbol{\Psi}^{(k+b/B)}$ denote the estimate of $\boldsymbol{\Psi}$ after the bth iteration on the $(k+1)$th scan ($b = 1, \ldots, B$) and S_{b+1} denote the subset of $\{1, \ldots, n_{PL}\}$ containing the subscripts of those pseudo-leaf nodes that belong to the $(b+1)$th block

$(b = 0, \ldots, B\text{-}1)$. Suppose that a set of A_m is selected on the kth scan for $m = 1, \ldots, n_{PL}$. That is, on the $(b+1)$th iteration of the kth scan $(b = 0, \ldots, B\text{-}1)$, if $\tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k-1+b/B)}) < 0.005$ for $m \in S_{b+1}$, then A_m contains the ith component; otherwise A_m^c (the complement of A_m) contains i. Now suppose that the SPIEM step is to be implemented on the subsequent B iterations of the $(k+1)$th scan. Then on the $(b+1)$th iteration $(b\ = 0, \ldots, B\text{-}1)$, consider for all $m \in S_{b+1}$,

- for all $i \in A_m$, set $\tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k+b/B)}) = \tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k-1+b/B)})$,
- for all $i \in A_m^c$, calculate the "non-proper" posterior probabilities of component membership, denoted as $\tau_i^*(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k+b/B)})$, based on the current estimates $\boldsymbol{\Psi}^{(k+b/B)}$ and then update posterior probabilities $\tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k+b/B)})$ by rescaling $\tau_i^*(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k+b/B)})$ as

$$\tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k+b/B)}) = \left[\sum_{h \in A_m^c} \tau_h(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k-1+b/B)}) \right] \frac{\tau_i^*(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k+b/B)})}{\sum_{h \in A_m^c} \tau_h^*(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k+b/B)})}.$$

This sparse version of the partial E-step thus will take time proportional only to the number of components $i \in A_m^c$ $(m = 1, \ldots, n_{PL})$. The current conditional expectations of the sufficient statistics $T_{i1}^{(k+\,b/B)}$, $\boldsymbol{T}_{i2}^{(k+\,b/B)}$, and $\boldsymbol{T}_{i3}^{(k+\,b/B)}$ are obtained for $i = 1, \ldots, g$, using the relationship

$$\boldsymbol{T}_{iq}^{(k+\,b/B)} = \boldsymbol{T}_{iq}^{(k+\,(b-1)/B)} - \boldsymbol{T}_{iq,b+1}^{(k-1+\,b/B)} + \boldsymbol{T}_{iq,b+1}^{(k+\,b/B)} \quad (q = 1, 2, 3) \qquad (8)$$

for $b = 0, \ldots, B\text{-}1$, where the first and the second terms on the right-hand side of (8) are available from the previous iteration and previous scan, respectively. Only the third term of (8) have to be calculated by updating only the contribution to the sufficient statistics for those components $i \in A_m^c$. For example,

$$\begin{aligned} T_{i1,b+1}^{(k+b/B)} = &\sum_{m \in S_{b+1}} I_{A_m}(i) \tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k-1+b/B)}) n_m \\ &+ \sum_{m \in S_{b+1}} I_{A_m^c}(i) \tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k+b/B)}) n_m, \end{aligned} \qquad (9)$$

where $I_{A_m}(i)$ is the indicator function for the set A_m. The first term on the right-hand side of (9) is calculated at the $(b+1)$th iteration of the kth scan and can be saved for use in the subsequent iteration on the $(k+1)$th scan. Similar arguments apply to $\boldsymbol{T}_{i2}^{(k+b/B)}$ and $\boldsymbol{T}_{i3}^{(k+b/B)}$.

3 Robust estimation

Robust fitting of Gaussian mixtures has been considered, using M-estimates to update the component estimates on the M-step of the EM algorithm [8, 9]. With the M-estimation on mrkd-trees, the updated component means $\boldsymbol{\mu}_i^{(k+1)}$ in (6) are replaced by

$$\boldsymbol{\mu}_i^{(k+1)} \approx \frac{\sum_{m=1}^{n_{PL}} \tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k)}) n_m u_{im}^{(k)} \bar{\boldsymbol{x}}_m}{\sum_{m=1}^{n_{PL}} \tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k)}) n_m u_{im}^{(k)}}, \qquad (10)$$

where $u_{im}^{(k)} = \psi(\Delta_{im}^{(k)})/\Delta_{im}^{(k)}$, and where $\Delta_{im}^{(k)} = \{(\bar{\boldsymbol{x}}_m - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\bar{\boldsymbol{x}}_m - \boldsymbol{\mu}_i)\}^{1/2}$ is the Mahalanobis distance between vectors $\bar{\boldsymbol{x}}_m$ and $\boldsymbol{\mu}_i$, and $\psi(s) = -\psi(-s)$ is Huber's [10] ψ-function defined as

$$\begin{aligned} \psi(s) &= s, & \mid s \mid \leq a, \\ &= sign(s)a, & \mid s \mid > a, \end{aligned} \tag{11}$$

for an appropriate choice of the tunning constant a. Similarly, the ith component-covariance matrix $\boldsymbol{\Sigma}_i^{(k+1)}$ in (7) is replaced by

$$\boldsymbol{\Sigma}_i^{(k+1)} \approx \frac{\sum_{m=1}^{n_{PL}} \tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k)}) n_m u_{im}^{2\ (k)} (\bar{\boldsymbol{x}}_m - \boldsymbol{\mu}_i^{(k+1)})(\bar{\boldsymbol{x}}_m - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{m=1}^{n_{PL}} \tau_i(\bar{\boldsymbol{x}}_m; \boldsymbol{\Psi}^{(k)}) n_m u_{im}^{2\ (k)}}. \tag{12}$$

An alternative to Huber's ψ-function is a redescending ψ-function where observations that are extremely atypical of a component will have zero weight for values of $\Delta_{im}^{(k)}$ above a certain level (rejection point) [2, Section 7.6]. A review on robust clustering methods using statistical approaches and fuzzy set theory can be found in [11].

With the mrkd-trees structure, we perform a robust estimation for Gaussian mixtures by identifying tree-nodes as three different types. Different weights are then given on them in the calculation of parameters. The computations involved in the categorization of tree-nodes can be readily obtained by using only the kd-tree code of the SPIEM mrkd-tree algorithm. Thus extra burden of computation is not required. The type of each tree node is determined at the pruning process and is based on the "denseness" of the node and the squared distance d_{im} between $\bar{\boldsymbol{x}}_m$ and the current estimated Gaussian centre $\boldsymbol{\mu}_i$ $(i = 1, \ldots, g; m = 1, \ldots, n_{PL})$.

The first type is that the node is close to at least one of g components. Let λ_i and λ_i' denote the smallest and the largest eigenvalues of $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$, which are, respectively, the minimum and the maximum values of the Mahalanobis squared distance for all points on unit sphere. For the mth node, if the squared distance

$$d_{hm} < \lambda_h \qquad \text{for some } h \in \{1, \ldots, g\},$$

then data points in this node are considered to be come from the main body (inlier) of Gaussian mixture. Full weight $u_{im} = 1$ is given to this node for all $i = 1, \ldots, g$. For calculating $\boldsymbol{T}_{iq}$ $(i = 1, \ldots, g; q = 1, 2, 3)$ in (8), we search the leaf nodes under the mth node for the hth component and prune the node for other components $i \neq h$. This improves the accuracy of the estimates.

The second type is that the node is far away from all the Gaussian centres and is not dense. The former condition is determined if

$$d_{im} > 4\lambda_i' \qquad \text{for all } i \ (i = 1, \ldots, g).$$

The latter is determined if (i) the number of data points in the mth node is smaller than 10, and (ii) the maximum diagonal element of the sample covariance matrix $\boldsymbol{S}_m$ of data points in the node, say in the vth dimension $v \in \{1, \ldots, p\}$, satisfies

$$(\boldsymbol{S}_m)_{vv} > 0.1(\boldsymbol{S})_{vv},$$

where $\boldsymbol{S}$ is the global sample covariance of the whole data set. Data points in this node are then considered to be come from the noise (outlier of Gaussian mixture) and reduced weight $u_{im} = 1/\Delta_{im}$ is given for all $i = 1, \ldots, g$. A dense node is not considered as an outlier automatically because a moderate size cluster of data points may not arise simply by chance (noise) and could be an interesting feature of the data required further investigation.

All nodes that are not identified as above two types form the third category. The weight given to these nodes is based on Huber's ψ-function (11), $u_{im} = \psi(\Delta_{im})/\Delta_{im}$, where $a^2 = \chi^2_{p,0.95}$ is adopted [9]. Thus, nodes that are atypical of a component are being given reduced weight in the calculation of parameters.

As an example, Fig. 2(a) shows a simulated data set of Gaussian mixture with noise. Nodes that are visited during the second scan of the proposed algorithm are categorized into three types (Fig. 2(b)). It is noted that large sized nodes (and hence large savings) are observed in areas with less variation in the component-posterior probabilities.

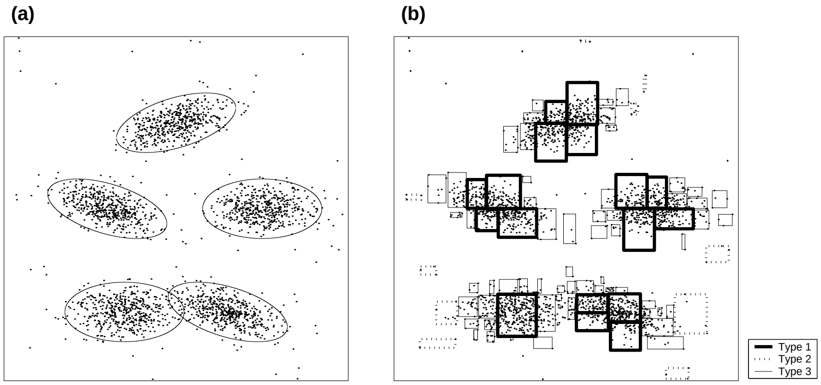

Fig. 2. Simulated Gaussian mixture with noise. Nodes are categorized into three types

4 Examples

Here we illustrate the proposed algorithm using simulated and real data. The simulated data consists initially of 50000 data points generated from a eight-component bivariate Gaussian mixtures, to which 5000 noise points were added from a uniform distribution over the range -10 to 10 on each variate. The parameters of the mixture model were

$$\boldsymbol{\mu}_1 = (3 \quad 0)^T, \quad \boldsymbol{\mu}_2 = (3 \quad -6)^T, \quad \boldsymbol{\mu}_3 = (-6 \quad 5)^T, \quad \boldsymbol{\mu}_4 = (5 \quad 7)^T,$$

$$\boldsymbol{\mu}_5 = (-4 \quad -6)^T, \quad \boldsymbol{\mu}_6 = (-1 \quad 7)^T, \quad \boldsymbol{\mu}_7 = (0 \quad 3)^T, \quad \boldsymbol{\mu}_8 = (-3 \quad 0)^T,$$

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}, \; \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 0.1 \end{pmatrix}, \; \boldsymbol{\Sigma}_4 = \begin{pmatrix} 1 & -0.1 \\ -0.1 & 0.1 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_5 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \ \boldsymbol{\Sigma}_6 = \begin{pmatrix} 2 & 0 \\ 0 & 0.5 \end{pmatrix}, \ \boldsymbol{\Sigma}_7 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \ \boldsymbol{\Sigma}_8 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{pmatrix},$$

with equal mixing proportions $\pi_i = 1/8$ $(i = 1, \ldots, 8)$. The true grouping of the eight-component Gaussian mixture is shown in Fig. 3(a). We now consider the clustering obtained by the robust estimation of the proposed SPIEM *mrk*d-tree algorithm. The clustering so obtained is given in Fig. 3(b). It compares well with the true grouping in Fig. 3(a). The result of fitting Gaussian mixture of eight components is given in Fig. 3(c) for comparison. It can be seen that the eight-component mixture fails to identify correctly some covariance matrices.

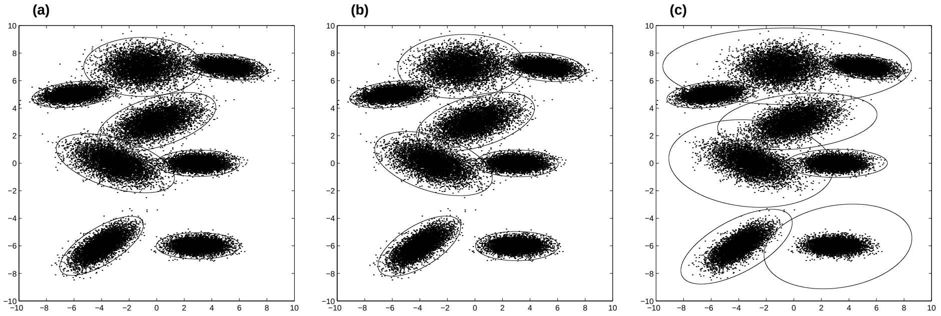

Fig. 3. Results for simulated Gaussian mixture with noise

A more complex mixture model may be adopted to model the additional background noise. If the number of components is treated as unknown and a Gaussian mixture is fitted, then the number of components selected via the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) is eleven [2, Sections 6.8–6.9]. The additional three components are attempting to model the background noise. However, estimation of some covariance matrices is still affected by the noise. In comparing the computational performance of these algorithms, the same initialization procedure was used in this simulation study. Ten trials of k-means with two scans were performed for each model to initialize the EM-based algorithms [2, pp. 98]. The number of scans and the CPU time (in seconds) required for various models are presented in Table 1.

Table 1. Computational performances for simulated Gaussian mixture with noise

Method	No. of scans	CPU time (sec.)
SPIEM–*k*d-tree	12	5
EM (8 components)	44	77
EM (11 components)	134	322

We now apply the proposed algorithm to segment a real 2D MR image data of the human brain into three regions (gray matter, white matter, and cerebrospinal fluid) in the presence of background noise arising from instrument irregularities. The data set was acquired by a two-Tesla Bruker Medspac whole body scanner. The acquisition matrix was 256 × 256. For each pixel, T_1-, T_2-, and ρ_D-weighted image intensities were available. Fig. 4(a) displays the T_1-weighted image. In the analysis, the volume of interest (VOI) was determined using a mask of head. This eliminates part of the surrounding tissues of the brain. A procedure for computing the mask is described in [12]. This step does not need to be precise as robust estimation will be undertaken subsequently. We considered $g = 3$ corresponding to the three main tissue types as described above. The result is displayed in Fig. 4(b). It can be seen that the three tissue types are well separated. It is noted that some dark spots, which correspond to the outlier, are observed. These can be corrected by applying a Markov random field (MRF) model to capture the spatial correlation in image intensities between neighboring pixels [5]. A "contextual" segmentation based on the MRF model of [5] (with parameter $\beta = 1$) is displayed in Fig. 4(c) for comparison.

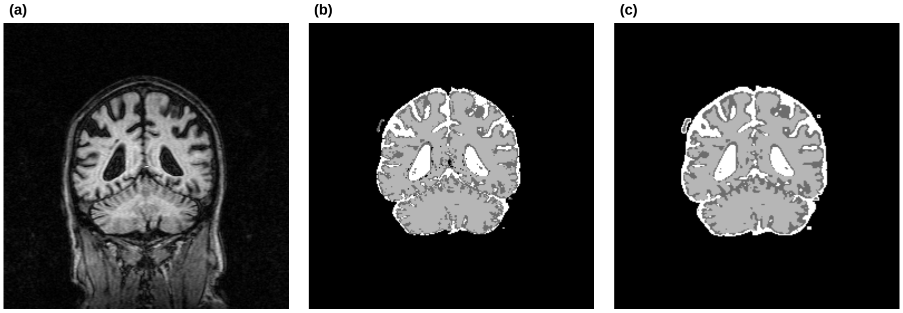

Fig. 4. Results for real 2D MR image data

5 Conclusions

We have described a robust version of the SPIEM *mrk*d-tree algorithm for speeding up the mixture model-based image segmentation. During the pruning process of *k*d-trees, nodes are categorized into three different types and different weights are then given in the calculation of parameters. The proposed method has been illustrated using real and simulated data. The *mrk*d-tree approach, however, will not be able to speed up the EM algorithm when the dimension of the data increases, for example, there appears to be little gain for $p \geq 6$. Dimensionality reduction methods may be adopted [4] although this is in general not a problem in the image segmentation as $p \leq 3$ usually holds. Alternative fast moment-based method has also been considered in numerical analysis [13].

In this paper, we focus on the robust estimation within the *k*d-trees framework. A detailed description of other robust approaches to the fitting of Gaussian mixtures can be found in [2, Chapter 7]. For example, the simulated data in Section 4 may also be well fitted by a eight-component Gaussian mixture and an additional uniform component. A similar model (a three-component univariate Gaussian mixture and a uniform distribution) has also been used to segment 1D MR images [14]. However, the model may not be work as well in situations when the noise is not uniform or is unable to be modeled adequately by the uniform distribution. In contrast, our method is able to speed up the segmentation process and at the same time provide robust estimation without much extra computational burden.

A drawback of our robust version is that the log likelihood values are no longer monotonically increasing after each scan (compared to [4]). The algorithm, however, can be terminated by considering the convergence of the estimates at each scan. Alternatively, the log likelihood value may be approximated by imposing weights u_{im} in its calculation. Weight functions in terms of the Pearson residuals have been considered in [15, 16]. The application of this weighted likelihood methodology within the kd-trees framework requires further investigation.

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B **39** (1977) 1–38
2. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
3. Moore, A.W.: Very fast EM-based mixture model clustering using multiresolution kd-trees. In: Kearns, M.S., Solla, S.A., Cohn, D.A. (eds.): Adv. in Neural Inf. Proc. Systems 11. MIT Press, Cambridge, Massachusetts (1999) 543–549
4. Ng, S.K., McLachlan, G.J.: On speeding up the EM algorithm in pattern recognition: a sparse version of the incremental EM algorithm based on a multiresolution kd-tree structure. Technical Report, Centre for Statistics, University of Queensland, Brisbane (2002)
5. McLachlan, G.J., Ng, S.K., Galloway, G., Wang, D.: Clustering of magnetic resonance images. Proc. Am. Stat. Assoc. (Stat. Comp. Sect.). Am. Stat. Assoc., Alexandria, Virginia (1996) 12–17
6. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of MR images of the brain. IEEE T. Med. Imaging **18** (1999) 897–908
7. Ng, S.K., McLachlan, G.J.: On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. Stat. Comput. **13** (2003) 45–55
8. Campbell, N.A.: Mixture models and atypical values. J. Int. Ass. Math. Geol. **16** (1984) 465–477
9. McLachlan, G.J., Basford, K.B.: Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York (1988) Section 2.8
10. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35** (1964) 73–101
11. Davé, R.N., Krishnapuram, R.: Robust clustering methods: a unified view. IEEE T. Fuzzy Syst. **5** (1997) 270–293
12. Brummer, M.E., Mersereau, R.M., Eisner, R.L., Lewine, R.R.J.: Automatic detection of brain contours in MRI data sets. IEEE T. Med. Imaging **12** (1993) 153–166
13. Beatson, R.K., Newsam, G.N.: Fast evaluation of radial basis functions: Moment based methods. SIAM J. Sci. Comput. **19** (1998) 1428–1449
14. Schroeter, P., Vesin, J.M., Langenberger, T., Meuli, R.: Robust parameter estimation of intensity distributions for brain magnetic resonance images. IEEE T. Med. Imaging **17** (1998) 172–186
15. Green, P.J.: Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. J. Roy. Stat. Soc. Ser. B **46** (1984) 149–192
16. Markatou, M., Basu, A., Lindsay, B.G.: Weighted likelihood equations with bootstrap root search. J. Am. Stat. Assoc. **93** (1998) 740–750