Published in final edited form as:
Comput Stat Data Anal. 2009 May 15; 53(7): 2563–2572. doi:10.1016/j.csda.2008.12.005.

Mixture modeling with applications in schizophrenia research

Qiang Wu[a,*] and **Allan R. Sampson**[b]

[a] Department of Biostatistics, East Carolina University, 2435D Health Sciences, Building, Greenville, NC, USA, 27858

[b] Department of Statistics, University of Pittsburgh, 2717 Cathedral of Learning, Pittsburgh, PA, USA, 15260

Abstract

Finite mixture modeling, together with the EM algorithm, have been widely used in clustering analysis. Under such methods, the unknown group membership is usually treated as missing data. When the "complete data" (log-)likelihood function does not have an explicit solution, the simplicity of the EM algorithm breaks down. Authors, including Rai and Matthews (1993), Lange (1995a) and Titterington (1984), developed modified algorithms therefore. As motivated by research in a large neurobiological project, we propose in this paper a new variant of such modifications and show that it is self-consistent. Moreover, simulations are conducted to demonstrate that the new variant converges faster than its predecessors.

Keywords

Clustering; EM algorithm; EM1 algorithm; EM-gradient algorithm; finite mixture models; schizophrenia

1 Introduction

Finite mixture modeling is a widely used clustering technique for moderate to highly complicated data structure (McLachlan and Basford, 1988). Such methods are model based and generate probabilities for the unknown group membership. The EM algorithm has been used in finding the parameter estimates for finite mixture models. It benefits in terms of stability and simplicity from the fact that the "complete data" (log-)likelihood function has an explicit solution. When the "complete data" (log-)likelihood function has no explicit solution, another iterative algorithm is needed in the M-step in order to maximize the conditional log-likelihood function given the observed data. This has been thought to be inefficient by many authors. Therefore, authors, including Rai and Matthews (1993), Lange (1995a) and Titterington (1984), proposed some modified algorithms which replaced the whole M-step of the EM algorithm with just one iteration of some Newton type algorithms in maximizing the conditional log-likelihood function. Given that the Newton type algorithms converges in a quadratic speed, the modified algorithms were shown to converge in a linear speed.

*Corresponding author. Tel.: +1 252 744 6047; fax: +1 252 744 6044., Email addresses: wuq@ecu.edu (Qiang Wu), asampson@pitt.edu (Allan R. Sampson).

On the other hand, when there are both outcome variables and covariates, different names, including “regression models for conditional normal mixtures” and “regression clustering”, have been given to the same problem. The basic idea is to cluster the subjects according to the discrepancy in the regression parameters or in addition the covariance parameters. DeSarbo and Corn (1988) defined a regression model for finite normal mixtures with a univariate outcome, while Jones and McLachlan (1992) extended the model to a multivariate setting. Arminger et al. (1999) introduced several likelihood based strategies for parameter estimation, including the EM algorithm and the EM-gradient algorithm (Lange, 1995a).

As motivated by research in a large neurobiological project, in this paper a new variant of the algorithms developed by Rai and Matthews (1993) and Titterington (1984) is proposed to a general finite mixture model. The main goal of our approach is to accelerate the algorithm without sacrificing its numerical simplicity. The rest of this paper is organized as follows. The motivation from the large neurobiological project is discussed in Section 2. Section 3 presents the derivation of the new algorithm. An application to the large neurobiological project is introduced in Section 4. Section 5 establishes the self-consistency of the new algorithm, whereas numerical simulations in investigating the speed of convergences is given in Section 6. Section 7 is the conclusions.

2 Motivation

Schizophrenia, a severe and disabling mental disease, has received a tremendous amount of research focus. A considerable amount of basic neurobiological research concerning schizophrenia have been conducted in the Conte Center for the Neuroscience of Mental Disorders in the Department of Psychiatry at the University of Pittsburgh. Differing neurobiological measurements from post-mortem brain tissue of subjects in the Brain Bank Core of the Center are involved in numerous studies. These studies typically attempt to identify neurobiological markers that are differentially expressed in subjects with schizophrenia as compared to normal controls (Konopaske et al., 2006). Consequently, it is of interest to attempt to undertake a large neurobiological project to integrate data from multiple Center’s studies. There is limited literature on previous attempts at such a data synthesis for post-mortem tissue studies in schizophrenia research. Two such papers are by Knable et al. (2001, 2002). The ultimate goal of our project is to identify possible heterogeneous groups of subjects with schizophrenia based on the various neurobiological markers, and this requires a series of major methodological steps. Clearly, the purpose of our long-term research is to provide new insights into the understanding of the neurobiology of schizophrenia.

The Conte Center databases we need to use involve numerous studies with varying subject populations and differing types of data. The main data issues include repeated measurements, differing matched controls for the same subject with schizophrenia and scientifically important covariates. Whenever repeated measurements occur, we plan to combine them into a single observation appropriate to that study. Multivariate normal models with structured means and covariance matrices have been developed by Wu and Sampson (2008) to deal with the differing matched controls and covariates.

Although the data are usually assumed to be normally distributed in the context of post-mortem tissue studies in schizophrenia, we derive our new procedure in a general setting with arbitrary component distributions. An application to the data structure in the large neurobiological project is illustrated as an example. This application uses the structured models introduced in Wu and Sampson (2008) which are revisited in Section 4. The actual data we will eventually combine from multiple studies show a considerable degree of missingness. To implement the clustering algorithm for the Center’s databases will require carefully crafted multiple

imputation schemes. And due to the high computational burden in multiple imputations, time efficiency of the corresponding complete data clustering algorithms is of concern.

3 Derivation of the new algorithm

For a sample of n independent multivariate observations $\mathbf{y}_1, \cdots, \mathbf{y}_n$, we consider a setting of finite mixture models with $g \geq 2$ subpopulations. Let $\mathbf{z}_i = (z_{i1}, \cdots, z_{ig})'$, $i = 1, \ldots, n$, be the unobserved group indicators, where $z_{ik} = 0$ or 1 and $\sum_{k=1}^{g} z_{ik} = 1$. And $z_{ik} = 1$ implies that $\mathbf{y}_i$ is a random sample from the kth subpopulation. Marginally, $\mathbf{z}_1, \cdots, \mathbf{z}_n$ are independent and identically distributed with a multinomial$(1; \pi_1, \cdots, \pi_g)$ distribution, where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{g} \pi_k = 1$. Conditional on $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$, we assume $\mathbf{y}_1, \cdots, \mathbf{y}_n$ having density functions given by

$$f(\mathbf{y}_i | z_{ik} = 1) = f_{ik}(\mathbf{y}_i; \boldsymbol{\theta})$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, g$, where $f_{ik}(\cdot; \cdot)$ depends on i through, e.g., some external covariates of the subject or the design of the study, and the subscript k implies the kth mixture component as identified by parts of $\boldsymbol{\theta}$. One example is $f_{ik}(\mathbf{y}_i; \boldsymbol{\theta}) = \varphi(\mathbf{y}_i; X_i\boldsymbol{\beta}_k, \Sigma_i(\boldsymbol{\sigma}))$, where $\varphi(\cdot; \boldsymbol{\mu}, \Sigma)$ denotes the multivariate normal density function with mean μ and covariance matrix Σ, X_i is the known covariates matrix, and $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \cdots, \boldsymbol{\beta}_g', \boldsymbol{\sigma}')'$ is the vector of unknown parameters. In this case, $f_{ik}(\mathbf{y}_i; \boldsymbol{\theta})$ depends on i through X_i and the design of $\Sigma_i(\boldsymbol{\sigma})$. Two possible designs of $\Sigma_i(\boldsymbol{\sigma})$ include dimension changing from subject to subject and random effect covariates. The parameter vector $\boldsymbol{\beta}_k$ identifies the kth mixture component. Here, the covariates X_i is formed as a matrix with each row corresponding to each component in $\mathbf{y}_i$. This formulation is different from the conventional one used in multivariate linear regressions where X_i is a vector and $\boldsymbol{\beta}_k$ is a matrix (Section 7.7, Johnson and Wichern, 2002). However, it provides more flexibility in modeling the mean structure of $\mathbf{y}_i$ in the way that each component of $\mathbf{y}_i$ can now have different covariates values. And it made the following notation easier.

In the above settings, the likelihood function for the observed data $Y = (\mathbf{y}_1, \cdots, \mathbf{y}_n)'$ is given by

$$L(\boldsymbol{\pi}, \boldsymbol{\theta} | Y) = \prod_{i=1}^{n} [\sum_{k=1}^{g} \pi_k f_{ik}(\mathbf{y}_i; \boldsymbol{\theta})] \tag{1}$$

where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_{g-1})'$ since π_g is redundant given $\sum_{k=1}^{g} \pi_k = 1$. However, directly maximizing (1) is intractable, even if $f_{ik}(\mathbf{y}_i; \boldsymbol{\theta})$ has a simple form. Instead, the EM algorithm introduced by Dempster, Laird, and Rubin (1977) focuses on the complete (augmented) data (Y, Z) and its likelihood function

$$L(\boldsymbol{\pi}, \boldsymbol{\theta} | Y, Z) = \prod_{i=1}^{n} \prod_{k=1}^{g} [\pi_k f_{ik}(\mathbf{y}_i; \boldsymbol{\theta})]^{z_{ik}}.$$

Denote $\boldsymbol{\vartheta} = (\pi', \boldsymbol{\theta})'$ for notational ease. The observed likelihood function (1) is then maximized though iterative maximizations of the conditional expectation

$$Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})=E_{\boldsymbol{\vartheta}^{(t)}}[\log L(\boldsymbol{\vartheta}|Y,Z)|Y]=\sum_{i=1}^{n}\sum_{k=1}^{g}\tau_{ik}^{(t)}[\log\pi_k+\log f_{ik}(\mathbf{y}_i;\boldsymbol{\theta})] \tag{2}$$

for $t = 0, 1, 2, \ldots$, where

$$\tau_{ik}^{(t)}=\frac{\pi_k^{(t)} f_{ik}(\mathbf{y}_i;\boldsymbol{\theta}^{(t)})}{\sum_{j=1}^{g}\pi_j^{(t)} f_{ij}(\mathbf{y}_i;\boldsymbol{\theta}^{(t)})}.$$

The EM algorithm consists of iterations of an E-step which computes (2) and an M-step which maximizes (2) with respect to its first argument. The EM algorithm takes advantage from the fact that

$$\frac{\partial Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial\boldsymbol{\vartheta}}=0 \tag{3}$$

has a closed form solution. If this is not the case, then the EM algorithm loses its base merit of simplicity. In fact, it is easy to see that there is always a closed from solution for $\boldsymbol{\pi}$ given by

$$\pi_k^{(t+1)}=\frac{\sum_{i=1}^{n}\tau_{ik}^{(t)}}{n} \tag{4}$$

for $k = 1, \ldots, g$.

When the closed form solution for $\boldsymbol{\theta}$ is not available, the EM algorithm requires another iterative algorithm, e.g., Newton-Raphson, in the M-step to solve (3). Many researchers find this second set of iterations to be inefficient. According to McLachlan and Krishnan (2008, p. 25 and pp. 149–153), Rai and Matthews (1993) proposed an EM1 algorithm where they replaced the entire M-step of the EM algorithm with one iteration of the Newton's method given by

$$\boldsymbol{\theta}^{(t+1)}=\boldsymbol{\theta}^{(t)}-\alpha^{(t)}\left[\frac{\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\vartheta}^{(t)}}\left[\frac{\partial Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\vartheta}^{(t)}}$$

where $0 < \alpha \leq 1$ was used to adjust the step size in each iteration to ensure the monotonicity of $L(\boldsymbol{\vartheta}^{(t)}|Y)$ with respect to t. A special version of the EM1 algorithm with $a^{(t)} \equiv 1$ was later referred to as the EM-gradient algorithm by Lange (1995a). In addition, Lange (1995a) considered choosing $a^{(t)} \equiv a$ to adjust the EM-gradient iterations and speed up convergence. This modification then served as a basis of a quasi-Newton acceleration introduced in Lange (1995b) where he noted that the EM-gradient actually acquired almost identical local convergence properties as the EM algorithm.

As another approach, Titterington (1984) proposed to update $\boldsymbol{\theta}$ according to

$$\boldsymbol{\theta}^{(t+1)}=\boldsymbol{\theta}^{(t)}-\alpha^{(t)}[I_c(\boldsymbol{\theta})]^{-1}_{\boldsymbol{\vartheta}^{(t)}}[\frac{\partial Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial\boldsymbol{\theta}}]_{\boldsymbol{\vartheta}^{(t)}}$$

where

$$I_c(\boldsymbol{\theta})=E_{\boldsymbol{\vartheta}}[\frac{\partial^2\log L(\boldsymbol{\vartheta}|Y,Z)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]$$

is the complete data information matrix with respect to $\boldsymbol{\theta}$. For a variety of models, e.g., mixtures with normal densities, $I_c(\boldsymbol{\theta})$ has a simpler form than $\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})/\partial\,\boldsymbol{\theta}\partial\,\boldsymbol{\theta}'$ in Rai and Matthews' and Lange's methods. In fact, it is not hard to show that under mild regularity conditions

$$[I_c(\boldsymbol{\theta})]_{\boldsymbol{\vartheta}^{(t)}}=E[\frac{\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]_{\boldsymbol{\vartheta}^{(t)}}. \tag{5}$$

Hence Titterington's algorithm can be thought of as a scoring version of the EM1 or EM-gradient algorithm. Similar as the EM1 algorithm, a fractional step achieved by a small enough $a^{(t)}$ will ensure the increase of $L(\boldsymbol{\vartheta}^{(t)}|Y)$.

Now let us take a detailed look at the matrices $\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})/\partial\,\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ and $I_c(\boldsymbol{\theta})$ given by

$$\frac{\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}=\sum_{i=1}^{n}\sum_{k=1}^{g}\tau_{ik}^{(t)}[\frac{\partial^2\log f_{ik}(\mathbf{y}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]$$

and

$$I_c(\boldsymbol{\theta})=\sum_{i=1}^{n}\sum_{k=1}^{g}\pi_k^{(t)}E[\frac{\partial^2\log f_{ik}(\mathbf{y}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}|z_{ik}=1]. \tag{6}$$

Titterington's algorithm is possibly superior to the EM1 algorithms in terms of their speed of convergence when $E[\partial^2\log f_{ik}(\mathbf{y}_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'|z_{ik}=1]$ is much simpler than $\partial^2\log f_{ik}(\mathbf{y}_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'$. But it also suffers from using the unconditional clustering probabilities $\pi_1^{(t)},\cdots,\pi_g^{(t)}$ in (6) instead of the conditional ones $\tau_{i1}^{(t)},\cdots,\tau_{ig}^{(t)}$, since it is clear that $\tau_{i1}^{(t)},\cdots,\tau_{ig}^{(t)}$ fit the data better than $\pi_1^{(t)},\cdots,\pi_g^{(t)}$ do in the current stage. Because $\{\tau_{ik}^{(t)}\}$ is necessarily computed for (4) in every iteration, we propose, as a variant to the above algorithms, to update $\boldsymbol{\theta}$ according to

$$\boldsymbol{\theta}^{(t+1)}=\boldsymbol{\theta}^{(t)}-\alpha^{(t)}[H(\boldsymbol{\theta})]^{-1}_{\boldsymbol{\vartheta}^{(t)}}[\frac{\partial Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial\boldsymbol{\theta}}]_{\boldsymbol{\vartheta}^{(t)}} \tag{7}$$

where

$$H(\boldsymbol{\theta})=\sum_{i=1}^{n}\sum_{k=1}^{g}\tau_{ik}^{(t)}E[\frac{\partial^2\log f_{ik}(\mathbf{y}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}|z_{ik}=1].$$

As can be seen, the new modification shares the same numerical simplicity as Titterington's algorithm. The self-consistency of the new algorithm is shown in Section 5. In the simulations in Section 6, the new algorithm with $\alpha^{(t)} \equiv 1$ is compared to the EM-gradient algorithm and Titterington's algorithm in terms of their convergence speed and clustering accuracy.

4 An example involving structured models

Although the new algorithm developed in Section 3 is applicable in more general settings, in this section we introduce a special example involving structured models because these models are critical in analyzing the motivating databases. And in addition, this example is suitable for demonstrating some basic advantages of the new algorithm in analyzing complex databases.

As noted by Wu and Sampson (2008), several special considerations arise when we attempt to combine data from multiple studies in the Conte Center. First, by December 31, 2005 about 35 separate post-mortem tissue studies have been conducted. These studies involve different neurobiological measurements, such as neuron counts, neuron somal volume and certain mRNA expression levels, on differing subject populations. As a result, the combined data would involve multiple responses and a considerable degree of missingness. Carefully crafted multiple imputation schemes will be required in dealing with missing data. In this paper, we develop an efficient complete data clustering algorithm which will be implemented later in conjunction with multiple imputations. Second, In order to control for both experimental and demographical variations every subject with schizophrenia has been matched with a normal control subject in each study based on their ages at death, gender and post-mortem intervals. And paired differences between measurements on the subjects with schizophrenia and the corresponding controls are typically obtained and treated as primary data in the original analysis. This convention is adopted here. Nevertheless, the matched controls for a subject with schizophrenia might be different in different studies. To be more explicit, let's consider S_{i1}, ⋯, S_{ip} and C_{i1}, ⋯, C_{ip} being p neurobiological measurements on, respectively, the ith subject with schizophrenia and its corresponding controls for $i = 1, 2, \cdots, n$. These measurements are obtained most possibly from more than one study. And there is a chance that C_{i1}, ⋯, C_{ip} are from different subjects. The pairwise differences are defined to be $S_{i1} - C_{i1}$, ⋯, $S_{ip} - C_{ip}$ for $i = 1, 2, \cdots, n$. Observations from the same subjects are assumed to be correlated, while observations from different subjects are taken to be independent. As a result, we have

$$Cov(S_{ij} - C_{ij}, S_{ik} - C_{ik}) = Cov(S_{ij}, S_{ik}) + Cov(C_{ij}, C_{ik}) \text{ for } j \neq k.$$

When C_{ij} and C_{ik} are from the same subject, $Cov(C_{ij}, C_{ik}) \neq 0$; otherwise, $Cov(C_{ij}, C_{ik}) = 0$. Finally, as common in studies with human subjects, covariates, such as age and gender, are involved in the Center's studies. The means of the multivariate responses $\mathbf{y}_i = (S_{i1} - C_{i1}, \cdots, S_{ip} - C_{ip})'$, $i = 1, 2, \cdots, n$, are then assumed to be linear in the covariates, that is

$$E[\mathbf{y}_i|X_i] = X_i\boldsymbol{\beta},$$

where X_i is the covariates matrix whose rows corresponding to the components in $\mathbf{y}_i$ and $\boldsymbol{\beta}$ is a vector of unknow parameters. This mean structure allows some covariates values to vary

from measurement to measurement. Examples of such covariates include tissue quality or solution density which are study dependent.

Wu and Sampson (2008) developed multivariate normal models with structured means and covariance matrices to deal with the special considerations in the Center's databases. The structured means resulted from the covariates of the subjects with schizophrenia, whereas the structured covariance matrices were due to the differing control subjects that were matched to the same subjects with schizophrenia when they were involved in different studies. Consider the responses $\mathbf{y}_1, \cdots, \mathbf{y}_n$ representing the pairwise differences between the n subjects with schizophrenia and their corresponding controls. Wu and Sampson (2008) defined $f_i(\mathbf{y}_i; \boldsymbol{\theta}) = \varphi(\mathbf{y}_i; X_i\boldsymbol{\beta}, \Sigma_i(\boldsymbol{\sigma}))$, $i = 1, \ldots, n$, as the corresponding density functions, where $X_1, \cdots, X_n$ are the known covariates matrices and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\sigma}')'$. The subscript i of $f_i(\mathbf{y}_i; \boldsymbol{\theta})$ denotes its dependency on i through X_i and $\Sigma_i(\boldsymbol{\sigma})$. Each $\Sigma_i(\boldsymbol{\sigma})$ was formulated to be linear in $\boldsymbol{\sigma}=(\sigma_{11},\cdots,\sigma_{pp},\sigma_{12},\cdots,\sigma_{(p-1)p},\sigma^c_{12},\cdots,\sigma^c_{(p-1)p})'$, i.e.,

$$\sum_i(\boldsymbol{\sigma})=\sum_{j=1}^{p}\sigma_{jj}G_{jj}+\sum_{1\le j<k\le p}(\sigma_{jk}+\sigma^c_{jk}I_i^{jk})G_{jk}$$

where G_{jk}, $1 \le j \le k \le p$, are known matrices with all "0" entries except a "1" at both the (j, k)th and the (k, j)th entries, and $I_i^{jk}=1$ if the controls matched to the ith subject with schizophrenia for the jth and the kth measurements are the same; otherwise, $I_i^{jk}=0$ (Wu and Sampson, 2008). Explicitly speaking, σ_{jj} is a sum of the jth measurement variances for both the subject with schizophrenia and the control, whereas σ_{jk} and σ^c_{jk} are the (j, k)th measurements covariances on the subject with schizophrenia and the control, respectively. And σ^c_{jk} is added onto the (j, k)th covariance of the pair-wise differences when $I_i^{jk}=1$, i.e., C_{ij} and C_{ik} are from the same subject. For data with p dimensional responses, there are a total of $2^p - p$ possible matching schemes between the subjects with schizophrenia and the controls. But most probably, not all of these matching schemes appear in one data set. In such cases, $I_i^{jk} \equiv 0$ or 1 for $i = 1, 2, \cdots, n$ for some $1 \le j < k \le p$. If $I_i^{jk} \equiv 0$, then σ_{jk} is estimable, but σ^c_{jk} is not; and If $I_i^{jk} \equiv 1$, then $\sigma_{jk}+\sigma^c_{jk}$ is estimable, but both summands are not. Thus, the number of parameters in $\boldsymbol{\sigma}$ needs to be reduced accordingly. However, in the following discussion we assume for simplicity that $0<\sum_{i=1}^{n} I_i^{jk}<n$ for all $1 \le j < k \le p$ so that all parameters in $\boldsymbol{\sigma}$ are estimable.

Following the model specification in Wu and Sampson (2008), in this example we consider a finite mixture model with component density functions $f_{ik}(\mathbf{y}_i; \boldsymbol{\theta}) = \varphi(\mathbf{y}_i; X_i\boldsymbol{\beta}_k, \Sigma_i(\boldsymbol{\sigma}))$ for $i = 1, \ldots, n$ and $k = 1, \ldots, g$, where $\boldsymbol{\theta}=(\boldsymbol{\beta}'_1,\cdots,\boldsymbol{\beta}'_g,\boldsymbol{\sigma}')'$. The subscript k of $f_{ik}(\mathbf{y}_i; \boldsymbol{\theta})$ identifies the kth mixture component. The purpose is to cluster the subjects with schizophrenia into possible subpopulations. In this model, the clusters are defined only in terms of the regression parameters $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_g$. There is no biological reason suggesting that the value of $\boldsymbol{\sigma}$ should be different over the clusters. For notational ease, we relabel the parameters in $\boldsymbol{\sigma}$ as $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_q)$. And by using some well-known matrix derivative results, we have

$$\frac{\partial Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial\boldsymbol{\beta}_k}=\sum_{i=1}^{n}\tau_{ik}^{(t)}X_i'\sum_i^{-1}(\mathbf{y}_i - X_i\boldsymbol{\beta}_k),\ \ k=1,\ldots,g, \tag{8}$$

$$\frac{\partial Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial \sigma_j}=\frac{1}{2}\sum_{i=1}^{n} tr\left(\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_j}\sum_i^{-1}\left(\sum_{k=1}^{g}\tau_{ik}^{(t)}C_{ik}-\sum_i\right)\right),\ \ j=1,\ldots,q \tag{9}$$

where $C_{ik} = (\mathbf{y}_i - X_i\boldsymbol{\beta}_k)(\mathbf{y}_i - X_i\boldsymbol{\beta}_k)'$. Continuing to take partial derivatives of (8) and (9) yields

$$\begin{aligned}
&-\frac{\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_k'}=\sum_{i=1}^{n}\tau_{ik}^{(t)}X_i'\sum_i^{-1}X_i,\ \ 1\le k\le g,\\
&-\frac{\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial \sigma_j \partial \sigma_l}=\frac{1}{2}\sum_{i=1}^{n} tr\left(\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_j}\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_l}i^{-1}\left(2\sum_{k=1}^{g}\tau_{ik}^{(t)}C_{ik}-\sum_i\right)\right)\\
&+\sum_{i=1}^{n} tr\left(\sum_i^{-1}\frac{\partial^2 \sum_i}{\partial \sigma_j \partial \sigma_l}\sum_i^{-1}\left(\sum_{k=1}^{g}\tau_{ik}^{(t)}C_{ik}-\sum_i\right)\right),\ \ 1\le j,l\le q,\\
&-\frac{\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k'}=0,\ \ 1\le k\ne j\le g,\\
&-\frac{\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial \boldsymbol{\beta}_k \partial \sigma_j}=\sum_{i=1}^{n}\tau_{ik}^{(t)}X_i'\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_j}\sum_i^{-1}(\mathbf{y}_i-X_i\boldsymbol{\beta}_k),\ \ 1\le k\le g,\ \ 1\le j\le q.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&-[I_c(\boldsymbol{\theta})]_{\boldsymbol{\beta}_k,\boldsymbol{\beta}_k}=\sum_{i=1}^{n}\pi_k^{(t)}X_i'\sum_i^{-1}X_i,\ \ 1\le k\le g,\\
&-[I_c(\boldsymbol{\theta})]_{\sigma_j,\sigma_l}=\frac{1}{2}\sum_{i=1}^{n} tr\left(\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_j}\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_l}\right),\ \ 1\le j,l\le q,\\
&-[I_c(\boldsymbol{\theta})]_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_k}=[I_c(\boldsymbol{\theta})]_{\boldsymbol{\beta}_k,\sigma_l}=0,\ \ 1\le k\ne j\le g,\ \ 1\le l\le q.
\end{aligned}$$

Thus, our newly proposed algorithm (7) with $\alpha^{(t)} \equiv 1$ yields an update given by

$$\boldsymbol{\beta}_k^{(t+1)}=\left[\sum_{i=1}^{n}\tau_{ik}^{(t)}X_i'\sum_i^{-1}X_i\right]_{\boldsymbol{\vartheta}^{(t)}}^{-1}\left[\sum_{i=1}^{n}\tau_{ik}^{(t)}X_i'\sum_i^{-1}\mathbf{y}_i\right]_{\boldsymbol{\vartheta}^{(t)}},\ \ k=1,\ldots,g, \tag{10}$$

$$\begin{aligned}
\boldsymbol{\sigma}^{(t+1)}=&\left[\sum_{i=1}^{n} tr\left(\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_j}\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_l}\right)\right]_{\boldsymbol{\vartheta}^{(t)}}^{-1}\left[\sum_{i=1}^{n} tr\left(\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_l}\sum_i^{-1}\left(\sum_{k=1}^{g}\tau_{ik}^{(t)}C_{ik}\right)\right)\right]_{\boldsymbol{\vartheta}^{(t)}}\\
&+\boldsymbol{\sigma}^{(t)}-\left[\sum_{i=1}^{n} tr\left(\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_j}\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_l}\right)\right]_{\boldsymbol{\vartheta}^{(t)}}^{-1}\left[\sum_{i=1}^{n} tr\left(\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_l}\right)\right]_{\boldsymbol{\vartheta}^{(t)}}.
\end{aligned} \tag{11}$$

When S_i is linear in the components of $\boldsymbol{\sigma}$ as described earlier, (11) becomes

$$\boldsymbol{\sigma}^{(t+1)}=\left[\sum_{i=1}^{n} tr\left(\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_j}\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_l}\right)\right]_{\boldsymbol{\vartheta}^{(t)}}^{-1}\left[\sum_{i=1}^{n} tr\left(\sum_i^{-1}\frac{\partial \sum_i}{\partial \sigma_l}\sum_i^{-1}\left(\sum_{k=1}^{g}\tau_{ik}^{(t)}C_{ik}\right)\right)\right]_{\boldsymbol{\vartheta}^{(t)}}. \tag{12}$$

Clearly, the quantities in (10) are the maximum likelihood estimates of the $\boldsymbol{\beta}$'s when the individual clustering probabilities and the covariance matrices are assume to be known. The

calculation of (12) is easier than that of the corresponding update in the EM-gradient algorithm and as easy as that of the corresponding update in Titterington's algorithm.

5 Local convergence properties

It was shown by Dempster, Laird, and Rubin (1977) that $L(\boldsymbol{\vartheta}^{(t+1)}|Y) =\geq L(\boldsymbol{\vartheta}^{(t)}|Y)$ as long as $Q(\boldsymbol{\vartheta}^{(t+1)}|\boldsymbol{\vartheta}^{(t)}) \geq Q(\boldsymbol{\vartheta}^{(t)}|\boldsymbol{\vartheta}^{(t)})$. So the convergence of the EM algorithm to a local maximum of $L(\boldsymbol{\vartheta}|\mathbf{y})$ is guaranteed under mild regularity condition. In addition, it can be shown that

$$[\frac{\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]_{\boldsymbol{\vartheta}^{(t)}} < [\frac{\partial^2 \log L(\boldsymbol{\vartheta}|Y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]_{\boldsymbol{\vartheta}^{(t)}}.$$

So the negative definiteness of $[\partial^2 Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})/\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}']_{\boldsymbol{\vartheta}^{(t)}}$ around a local maximum follows from the negative definiteness of $[\partial^2 \log L(\boldsymbol{\vartheta}|Y)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']_{\boldsymbol{\vartheta}^{(t)}}$. The negative definiteness of $[I_c(\boldsymbol{\theta})]_{\boldsymbol{\vartheta}^{(t)}}$ around a local maximum is also trivial. As a result, both Titterington's and the EM1 algorithms are necessarily ascent in the neighborhood of a local maximum, which means that there always exist at least one $0 < \alpha^{(t)} \leq 1$ which leads to an increase in $Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})$.

The same ascent property can also be established for our newly proposed algorithm. First, the negative definiteness of $[H(\boldsymbol{\theta})]_{\boldsymbol{\vartheta}^{(t)}}$ is guaranteed, since every individual item

$$E[\frac{\partial^2 \log f_{ik}(\mathbf{y}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}|z_{ik}=1]_{\boldsymbol{\vartheta}^{(t)}}$$

is negative definite. Here we can focus only on the parameter $\boldsymbol{\theta}$, because (4), as a solution to (3), automatically leads to an increase in $Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})$ by the GEM theory. As a result, given the values of $\pi_1^{(t)}, \cdots, \pi_g^{(t)}$, we rewrite $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ as $Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)})$ and consider the difference $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$. A Taylor's expansion around $\boldsymbol{\theta}^{(t)}$ is given as

$$\begin{aligned} &Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = [\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}}]'_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}) \\ &+\tfrac{1}{2}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})'[\frac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}), \end{aligned} \tag{13}$$

where $\boldsymbol{\theta}^*$ is between $\boldsymbol{\theta}^{(t+1)}$ and $\boldsymbol{\theta}^{(t)}$. Now plugging our new update (7) into (13), we have

$$\begin{aligned} &Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = -\alpha^{(t)}[\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}}]'_{\boldsymbol{\theta}^{(t)}}[H(\boldsymbol{\theta})]^{-1}_{\boldsymbol{\theta}^{(t)}}[\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}}]_{\boldsymbol{\theta}^{(t)}} \\ &+\tfrac{1}{2}(\alpha^{(t)})^2[\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}}]'_{\boldsymbol{\theta}^{(t)}}[H(\boldsymbol{\theta})]^{-1}_{\boldsymbol{\theta}^{(t)}}[\frac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]_{\boldsymbol{\theta}^*}[H(\boldsymbol{\theta})]^{-1}_{\boldsymbol{\theta}^{(t)}}[\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}}]_{\boldsymbol{\theta}^{(t)}} \\ &=[\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}}]'_{\boldsymbol{\theta}^{(t)}}(\tfrac{1}{2}(\alpha^{(t)})^2[H(\boldsymbol{\theta})]^{-1}_{\boldsymbol{\theta}^{(t)}}[\frac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]_{\boldsymbol{\theta}^*}[H(\boldsymbol{\theta}^{(t)})]^{-1}_{\boldsymbol{\theta}^{(t)}} \\ &-\alpha^{(t)}[H(\boldsymbol{\theta})]^{-1}_{\boldsymbol{\theta}^{(t)}})[\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}}]_{\boldsymbol{\theta}^{(t)}}. \end{aligned}$$

The above quantity is in the quadratic form. In order to have $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = 0$, it is sufficient that

$$\frac{1}{2}(\alpha^{(t)})^2[H(\boldsymbol{\theta})]^{-1}_{\boldsymbol{\theta}^{(t)}}[\frac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]_{\boldsymbol{\theta}^*}[H(\boldsymbol{\theta})]^{-1}_{\boldsymbol{\theta}^{(t)}} - \alpha^{(t)}[H(\boldsymbol{\theta})]^{-1}_{\boldsymbol{\theta}^{(t)}} \geq 0. \tag{14}$$

Since both $[H(\boldsymbol{\theta})]^{(t)}_{\boldsymbol{\theta}}$ and $[\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']^{*}_{\boldsymbol{\theta}}$ are negative definite matrices. Inequality (14) can be reduced to

$$\alpha^{(t)} I \leq [-\frac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]^{-1/2}_{\boldsymbol{\theta}^*}(-2[H(\boldsymbol{\theta})]_{\boldsymbol{\theta}^{(t)}})[-\frac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]^{-1/2}_{\boldsymbol{\theta}^*} \tag{15}$$

where I is the identity matrix with the same dimension as $H(\boldsymbol{\theta})$. It is not hard to show that inequality (15) is satisfied when $\alpha^{(t)}$ is chosen to be smaller than the smallest eigenvalue of the matrix on the right hand side of (15). Consequently, the value of $a^{(t)}$ selected such way always lead to an increase in $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ and so in $L(\boldsymbol{\vartheta}|Y)$.

As another usage, $\alpha^{(t)}$ can be adjusted to ensure that the parameter estimates fall in the parameter space. This is usually called step-halving. For example, one concern is that the estimated covariance matrices should be positive definite. By using a small $\alpha^{(t)}$ in (7), we are actually shrinking $\boldsymbol{\vartheta}^{(t+1)}$ toward $\boldsymbol{\vartheta}^{(t)}$. And usually the parameter space is an open set. So given that $\boldsymbol{\vartheta}^{(t)}$ is in the parameter space, there will always be an $\alpha^{(t)}$ small enough to guarantee that $\boldsymbol{\vartheta}^{(t+1)}$ will also be in the parameter space. And the direction of the inequality in (15) enables us to apply step-halving while guaranteeing an increase in $L(\boldsymbol{\vartheta}|Y)$. However, using a small $\alpha^{(t)}$ reduces the speed of convergence.

6 Simulations

Carefully crafted multiple imputation schemes will be required in the last step of our long term project when we attempt clustering in the presence of the substantial amount of missing data in the database from the Center's studies. The computational speed of the clustering component is critical in the long term project and we intend to use our new algorithm for this purpose. In this section, we show using simulations that our algorithm is computationally faster than both Titterington's and the EM-gradient algorithms. Data are simulated to be complete except the usual missing clustering indicators. A fixed value of $\alpha^{(t)} = 1$ is used. All algorithms are coded and running in R language.

Data with a structure conforming to the one described in Section 4 are simulated for the clustering analysis. The dimension of the outcomes $y_1, \cdots, y_n$ is assumed to be three. According to the component density functions

$$f_{ik}(\mathbf{y}_i, \boldsymbol{\theta}) = \varphi(\mathbf{y}_i; X_i\boldsymbol{\beta}_k, \sum\nolimits_i(\boldsymbol{\sigma}))$$

for $i = 1, \ldots, n$ and $k = 1, 2$, data sets containing data from two clusters are simulated. Each data set contains $n = 500$ subjects with 250 for each cluster. The covariates matrix X_i is in the form of

$$X_i=\begin{bmatrix} 1 & x_1 & x_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_1 & x_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & x_1 & x_2 \end{bmatrix}$$

for $i = 1, \ldots, n$. The constant '1' in the X_i's represents the overall difference between the subjects with schizophrenia and the controls. The values of x_1 are integers sampled uniformly from 20 to 80 to mimic the covariate age. And the values of x_2 are sampled uniformly from binary {0, 1} to simulate the covariate gender. This design of X_i allows the covariates effects with respect to the three components of y_i to be different. No study dependent covariate is considered in this simulation study. The two clusters differ only in the parameters for the mean structures, and let

$$\begin{aligned}\boldsymbol{\beta}_1&=(\beta^1_{11},\beta^1_{12},\beta^1_{13},\beta^1_{21},\beta^1_{22},\beta^1_{23},\beta^1_{31},\beta^1_{32},\beta^1_{33})'\\&=(-100, 2, 50, -50, 2, 50, -50, 1, 50)'\\\boldsymbol{\beta}_2&=(\beta^2_{11},\beta^2_{12},\beta^2_{13},\beta^2_{21},\beta^2_{22},\beta^2_{23},\beta^2_{31},\beta^2_{32},\beta^2_{33})'\\&=(100, -2, 50, 50, 2, 50, 50, -1, 50)'\end{aligned}$$

be those parameters for the two clusters, respectively. Explicitly speaking, $(\beta^1_{j1},\beta^1_{j2},\beta^1_{j3})'$ are the regression parameters for the jth component of y_i in the first cluster for $j = 1, 2, 3$, whereas $(\beta^2_{j1},\beta^2_{j2},\beta^2_{j3})'$ are the regression parameters for the jth component of y_i in the second cluster. In addition, let

$$\begin{aligned}\boldsymbol{\sigma}&=(\sigma_{11},\sigma_{22},\sigma_{33},\sigma_{12},\sigma_{13},\sigma_{23},\sigma^c_{12},\sigma^c_{13},\sigma^c_{23})'\\&=(1000, 1500, 1000, 400, 500, 600, 200, -100, -200)'\end{aligned}$$

which creates five possible individual covariance matrices for the outcomes as follows:

$$\sum_{(1)}=\begin{bmatrix} 1000 & 400 & 500 \\ 400 & 1500 & 600 \\ 500 & 600 & 1000 \end{bmatrix} \sum_{(2)}=\begin{bmatrix} 1000 & 600 & 500 \\ 600 & 1500 & 600 \\ 500 & 600 & 1000 \end{bmatrix} \sum_{(3)}=\begin{bmatrix} 1000 & 400 & 400 \\ 400 & 1500 & 600 \\ 400 & 600 & 1000 \end{bmatrix}$$
$$\sum_{(4)}=\begin{bmatrix} 1000 & 400 & 500 \\ 400 & 1500 & 400 \\ 500 & 400 & 1000 \end{bmatrix} \sum_{(5)}=\begin{bmatrix} 1000 & 600 & 400 \\ 600 & 1500 & 400 \\ 400 & 400 & 1000 \end{bmatrix}.$$

In this simulation study, we assume that the five possible matching schemes between subjects with schizophrenia and controls are equally likely. As a result, within each data set and each cluster, each of the above covariance matrices appears to 50 subjects. Although this assumption will most likely be violated in reality, this algorithm can be implemented in the same way.

Direct applications of both the EM-gradient algorithm and Titterington's algorithm to our simulated data is time consuming. The three algorithms are implemented on a random selection of 30 out of 500 simulated data sets and shown to provide the same parameter estimates. As a result, only our new algorithm is used for the parameter estimation for the rest 470 simulated data sets.

First, we examine in detail their computational speed. Each of the three algorithms is then implemented from the same starting values to find the parameter estimates. For the feasibility of comparison, the three algorithms are stopped according to the same criterion, that is, when

the change in the parameter estimates does not exceed a pre-defined limit. We observe that when the algorithms are started from near the true parameter values, they converge in almost the same number of steps. However, when the algorithms are started far from the true parameter values, they behave differently. A typical result from one of the thirty selected data sets is shown in Figure 1. The x-axis represents the number of iterations, while the y-axis represents the value of the observed log-likelihood function evaluated at the parameter estimates in each iteration. Some beginning iteration history for the three algorithms with low values (large negative values) of the log-likelihood function is not shown in Figure 1 for the feasibility of comparison. It can be seen that the EM-gradient algorithm converges in more than 80 steps, while Titterington's algorithm converges in about 65 steps. However, our new algorithm only requires about 25 steps to converge, which is a big advantage as compared to the other two. In addition to the results on the numbers of iterations before convergence, we also observe that the average cost of time per iteration is 4.67 seconds for the EM-gradient algorithm and about 0.7 seconds for both Titterington's algorithm and our new algorithm. However, we do recognize that the average cost of time per iteration does depend on the coding of the algorithms, the computing software and the hardware configuration. From Figure 1, the main feature of the new algorithm is that it requires significantly fewer iterations in finding the region containing a maximum when started far from the optimum, while its number of steps for subsequent "local refinement" is actually comparable to the two existing algorithms. In addition, we show in Table 1 the average numbers of steps to convergence and the mean per iteration time, as well as their corresponding standard deviations, for the three algorithms over the thirty selected data sets. The result confirms the above conclusion that our new algorithm is computational more effective than both the EM-gradient and the Titterington's algorithms.

For our current simulations, two different approaches for starting points are used for the purpose of demonstration. One approach is to choose parameters close to the true parameter values, and the other approach is by starting the algorithm from randomly generated clustering indices, i.e., a random starting point. For any single simulated data set, we define the final clustering result to be "successful" if the algorithm clusters more than 95% of its subjects correctly. For the 500 simulated data sets, we observe that by starting from near the true parameter values we get "successful" clustering results on 100% of the simulated data sets, while by starting from random point we obtain "successful" clustering results for about 95% of the simulated data sets. For the other 5% of the simulated data sets, the algorithm either does not converge (1.4%) or converges (3.6%) to a solution resulting in a clustering in which the subjects are clustered complete randomly. For those data sets with "successful" clustering results when starting from random clustering indices, we summarize the results of the parameter estimation in Table 2 as compared to the true parameter values. It can been seen that the parameter estimation is reasonably accurate when the algorithm finds the correct clusters. In fact, these results are surprisingly good. Typically, no one relies on one random starting point if one has no information about where to start. In order to increase the chance of identifying the correct clustering, we could always start the algorithm from multiple starting points and pick the solution maximizing the likelihood function as the result.

7 Conclusions

In this paper, some special features of finite mixture models, as compared to general missing data problems, are utilized to speed up the parameter estimation algorithms. The EM-gradient algorithm provides updates which fit the data better in each iteration, while Titterington's algorithm requires less calculation in each iteration. Our new algorithm takes their both advantages and is shown to acquire their nice local convergence properties as a heritage. In addition, we show by simulations that our new algorithm converges in fewer iterations than its two predecessors while providing the same parameter estimates. And the cost of time per iteration of our new algorithm should be comparable to that of Titterington's algorithm and

lower than that of the EM-gradient algorithm. As discussed, the database of the Center's studies which our ultimated project will use involves a great degree of missing data in addition to the unobserved clustering indicators. The next steps in our long term goal of clustering subjects with schizophrenia are to develop specially crafted multiple imputation techniques, implement the newly developed clustering algorithm to the multiply imputed data sets, and finally integrate the multiple clustering results to a single clustering result.

Acknowledgments

We thank Dr. David Lewis for his insights and allowing us use the data obtained in his lab. We also thank the Center researchers, especially Dr. Takanori Hashimoto, for their generous help. This research was supported by NIMH Grant 5P50MH045156-18 and P50 MH084053-01. We finally thank the referees for their insightful comments.

References

Arminger G, Stein P, Wittenberg J. Mixtures of conditional mean-and covariance-structure models. Psychometrika 1999;64 (4):475–494.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B 1977;39 (1):1–38.

DeSarbo WS, Corn LW. A maximum likelihood methodology for clusterwise linear regression. Journal of Classification 1988;5:249–282.

Johnson, RA.; Wichern, DW. Applied multivariate statistical analysis. 5. Prentice-Hall, Inc.; Upper Saddle River, NJ: 2002.

Jones PN, McLachlan GJ. Fitting finite mixture models in a regression context. Australian Journal of Statistics 1992;34 (2):233–240.

Knable MB, Bacrci BM, Barko JJ, Webster MJ, Torrey EF. Molecular abnormalities in the major psychiatric illnesses: Classification and regression tree (crt) analysis of post-mortem prefrontal markers. Molecular Psychiatry 2002;7:392–404. [PubMed: 11986983]

Knable MB, Torrey EF, Webster MJ, Bartko JJ. Multivariate analysis of prefrontal cortical data from the stanley foundation neuropathology consortium. Brain Research Bulletin 2001;55 (5):651–659. [PubMed: 11576762]

Konopaske GT, Sweet RA, Wu Q, Sampson AR, Lewis DA. Regional specificity of chandelier neuron axon terminal alterations in schizophrenia. Neuroscience 2006;138(1)

Lange K. A gradient algorithm locally equivalent to the em algorithm. Journal of the Royal Statistical Society B 1995a;57 (2):425–437.

Lange K. A quasi-newton acceleration of the em algorithm. Statistica Sinica 1995b;5:1–18.

McLachlan, GJ.; Basford, KE. Mixture models: inference and applications to clustering. Dekker; New York: 1988.

McLachlan, GJ.; Krishnan, T. The EM Algorithm and Extensions. 2. Wiley; New York: 2008.

Rai SN, Matthews DE. Improving the em algorithm. Biometrics 1993;49:587–591.

Titterington DM. Recursive parameter estimation using incomplete data. Journal of the Royal Statistical Society B 1984;46:257–267.

Wu, Q.; Sampson, AR. Structured modeling with applications to integrative analysis of post-mortem tissue studies in schizophrenia, unpublished. Dec. 2008 URL http://personal.ecu.edu/wuq/media/structuredmodel.pdf

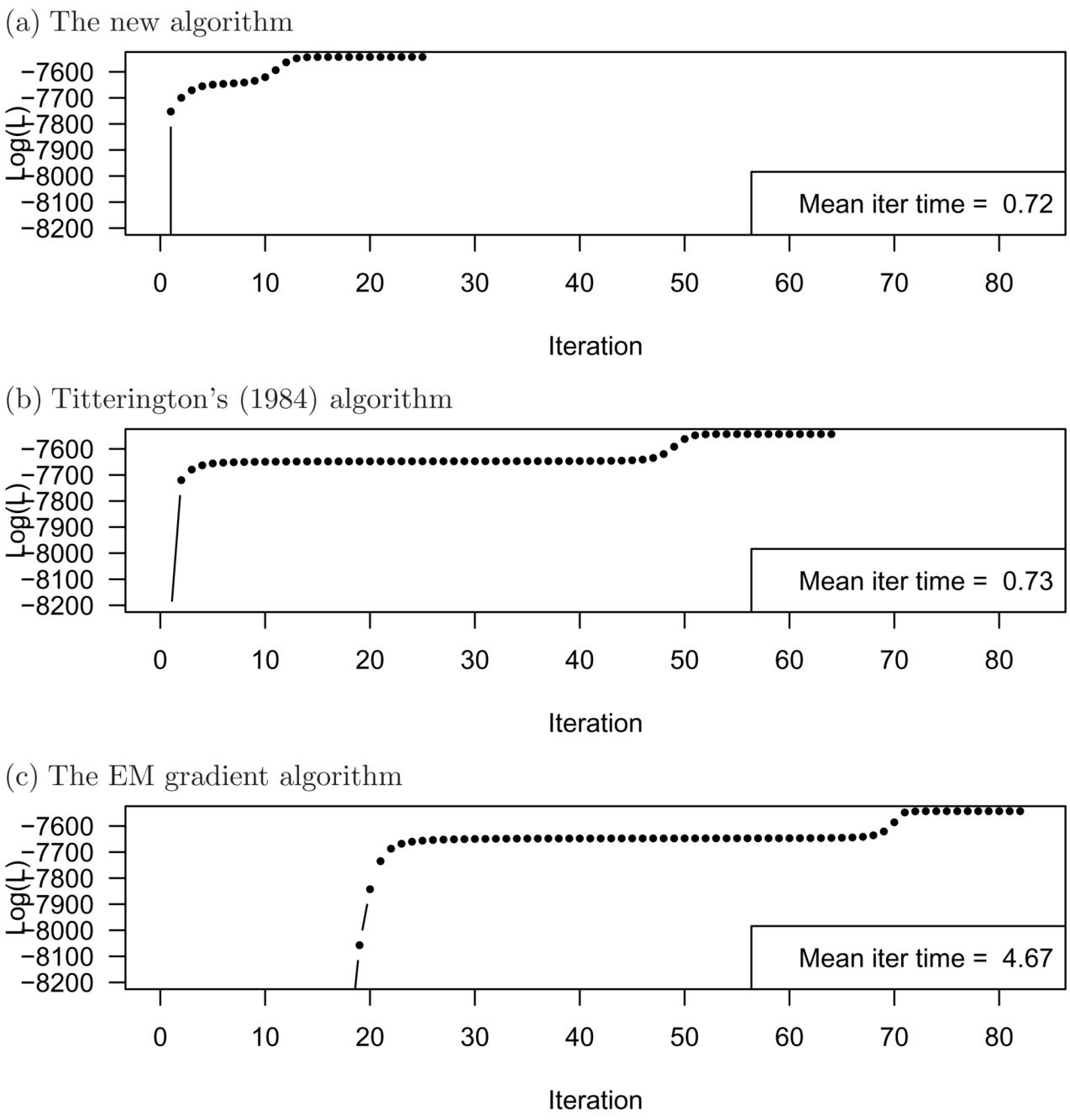

Fig. 1.
Iteration history of the clustering algorithms: (a) for the new algorithm; (b) for Titterington's (1984) algorithm; (c) for the EM gradient algorithm. The mean per iteration time, computed as the average of the computational time required for one iteration, is also shown.

Table 1

Average numbers of steps to convergence and mean per iteration time for the thirty simulated data sets

Algorithm	Average number of steps	std.	Average time per iteration	std.
New	41.1	15.2	0.728 sec	0.012 sec
Titterington	71.2	42.1	0.736 sec	0.008 sec
EM-gradient	75.4	31.1	4.583 sec	0.048 sec

Table 2

A Summary of the parameter estimates of the clustering simulations when the true clusters are identified

Param.	π	β_{11}^1	β_{12}^1	β_{13}^1	β_{21}^1	β_{22}^1	β_{23}^1	β_{31}^1	β_{32}^1	β_{33}^1
Truth [a]	0.5	−100	2	50	−50	2	50	−50	1	50
Mean [b]	0.50	−100.9	2.01	50.0	−50.5	2.02	49.1	−50.3	1.01	49.8
Std.[c]	0.01	6.47	0.11	4.10	7.67	0.14	5.25	6.33	0.11	4.16
Param.		β_{11}^2	β_{12}^2	β_{13}^2	β_{21}^2	β_{22}^2	β_{23}^2	β_{31}^2	β_{32}^2	β_{33}^2
Truth		100	−2	50	50	2	50	50	−1	50
Mean		99.7	−2.00	49.9	50.0	2.01	48.8	49.9	−1.00	50.1
Std.		6.74	0.12	4.21	8.00	0.15	4.97	6.80	0.12	4.15
Param.		σ_{11}	σ_{22}	σ_{33}	σ_{12}	σ_{13}	σ_{23}	σ_{12}^c	σ_{13}^c	σ_{23}^c
Truth		1000	1500	1000	400	500	600	200	−100	−200
Mean		1006.58	1502.33	971.43	458.32	483.26	544.26	160.24	−97.69	−155.43
Std.		69.85	127.95	64.09	69.85	56.52	69.72	71.45	64.06	78.00

[a] The true parameter values

[b] Means of the simulation estimates

[c] Standard deviations of the simulation estimates