Published in final edited form as:
Commun Stat Simul Comput. 2015 July ; 44(6): 1545–1556. doi:10.1080/03610918.2013.824091.

EM Estimation for Finite Mixture Models with Known Mixture Component Size

CHEN TEEL[1], **TAEYOUNG PARK**[2], and **ALLAN R. SAMPSON**[3]

[1]Applied Statistics Group, E. I. du Pont de Nemours & Company, DE, USA

[2]Department of Applied Statistics, Yonsei University, Seoul, Korea

[3]Department of Statistics, University of Pittsburgh, PA, USA

Abstract

We consider the use of an EM algorithm for fitting finite mixture models when mixture component size is known. This situation can occur in a number of settings, where individual membership is unknown but aggregate membership is known. When the mixture component size, i.e., the aggregate mixture component membership, is known, it is common practice to treat only the mixing probability as known. This approach does not, however, entirely account for the fact that the number of observations within each mixture component is known, which may result in artificially incorrect estimates of parameters. By fully capitalizing on the available information, the proposed EM algorithm shows robustness to the choice of starting values and exhibits numerically stable convergence properties.

Keywords

Aggregate data; Conditional Bernoulli distribution; EM algorithm; Finite mixture models

1 Introduction

Finite mixture models have been powerful tools for analyzing data where observations originate from various components but the component membership of each observation is not known. The analysis of such finite mixture models is commonly carried out using maximum likelihood estimation with the EM algorithm (Dempster et al., 1977; McLachlan and Peel, 2000). The finite mixture model has been extended to data with known mixture component size, i.e., the case when the exact number of observations within each mixture component is given and thus mixing probability parameters are known. The computational difficulty of fitting such a finite mixture model with known mixture component size, however, has hindered correct EM estimation and given rise to inconsistent EM estimates depending on starting values (Nettleton, 1999; Friede and Kieser, 2002). Here we focus on the problem of fitting a two-component mixture model with known mixture component size.

[0]Address correspondence to Taeyoung Park, Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea; tpark@yonsei.ac.kr.

The finite mixture problems with more than two mixture components would be solved by a nontrivial generalization of the method proposed in this paper.

Our work was specifically motivated by sample size adaptive designs for clinical trials. We want to consider adaptive designs that allow the estimation of nuisance parameters at an interim stage of a trial without breaking the blind for treatment identity. A common framework for studying such designs is to consider two treatments, experimental (E) and control (C) groups, with primary responses assumed to have normal distributions with common standard deviation σ and respective means μ_E and μ_C. Clinical trial sample sizes in such settings are typically based on power at a presumed value for σ and desired alternative $\mu_E - \mu_C$. In adaptive designs using sample size re-estimation, σ is suitably estimated in the actual trial from interim data obtained at the end of a pre-specified first phase and then an appropriate new sample size is usually calculated to maintain power. Various approaches have been considered to estimate σ from interim data for such adaptive designs (Zucker, 1999; Friede and Kieser, 2001; Xing, 2005). In particular, Gould and Shih (1992) proposed using the EM algorithm for the problem of estimating σ by fitting a two-component normal mixture model with known treatment group size for bimodal response data. To deal with known treatment group size, their EM algorithm fixed the mixing probability parameter at the proportion of patients assigned to the experimental group. By doing so, the EM algorithm devised by Gould and Shih (1992) sometimes produced an incorrect estimate toward the boundary of the parameter space (Nettleton, 1999).

In clinical trials, it is also common practice to randomize patients to treatments using randomly permuted blocks (Rosenberger and Lachin, 2002). Then, the allocation of patients to the experimental or control group is periodically “balanced” in such an adaptive design. This balancing is important when “time confounding” needs to be guarded against, especially for a long duration clinical trial. Time confounding can occur when medical equipment, concomitant medications, staff, and even disease severity of patients entering the trial at different times changes and patients are not allocated to the two treatments in some sort of balanced manner. Gould and Shih (1992) did not assume blocked randomization in doing the computation for their adaptive design (although they conceptually considered designs with planned equal number of patients at the interim for each treatment group but could use this for their EM procedure). Hence their EM algorithm used the planned proportion of numbers of patients to each treatment, e.g., 0.5, at the interim, so that the treatment identities are assumed to follow independent Bernoulli distributions with probability 0.5. Obviously this assumption does not guarantee an equal number of patients in each group at the interim.

Gould and Shih’s ideas and our awareness of block randomization led to developing our new EM algorithm for fitting finite mixture models with known mixture component size. We point out that in addition to knowing patient numbers at the interim, we can gain even more information by utilizing the individual randomization blocks prior to and including the interim. The full goal is to improve the estimation of σ from interim data by fully taking advantage of the information contained in the block randomization, yet maintaining the blind of the clinical trial.

In addition to adaptive clinical trials, our work is applicable to a number of other settings, where one knows that among n sample observations, there are m observations from one component and $n–m$ observations from the other, but which specific observations are from each component is not known. When the latent component membership is of interest as a function of observable covariates for each individual, such type of data have been extensively analyzed and growing in popularity with recent advances in computing; see Chen and Yang (2007); Choi et al. (2008); Musalem et al. (2009); Park (2011); Verhelst (2008) for various examples. Our goal is to estimate the underlying and unknown parameters that determine the distribution of each component type, where we allow more general distributions than the normal distribution. Another more specific type of application of our methodology is to voting inferences where we could make use of the fact that during an election between two candidates, we can obtain the exact number of votes each candidate receives at a voting site. Further, under certain conditions we can obtain the previous voting histories for each individual voter at that site. Because each voter's selection is blinded, the distribution of previous voting frequencies for individual voters can be viewed as a mixture of two distributions, each for two candidates' supporters. Thus, if we wanted to assess if there were a difference in previous voting patterns between those who voted for one candidate and those who voted for the other, we could use our methodology with m voters for the one candidate and $n – m$ for the other. In summary we see that the possible conceptual applications for our method arise when it is of interest to compare certain characteristics between two groups, given an anonymized list of two groups of people and the number of people belonging to each group.

In Section 2, we develop the EM algorithm for fitting mixture models of exponential family distributions when the exact number of observations within each mixture component is fixed. Section 3 discusses the efficient and numerically stable computation of the proposed EM algorithm. Section 4 compares in the context of normal mixture models the properties of the proposed EM algorithm and a conventional EM algorithm which uses the probability of the mixture being m/n. As an application of the proposed EM algorithm, a realistic adaptive design clinical trial example appears in Section 5. A discussion follows in Section 6.

2 The EM Algorithm with Known Mixture Component Size

2.1 The Conventional EM Algorithm

Suppose that $\mathbf{y} = (y_1, \ldots, y_n)$ are observations from a mixture of two exponential family distributions of the same type with common parameter ξ and respective parameters ψ_1 and ψ_2, collectively denoted by $\boldsymbol{\theta} = (\xi, \psi_1, \psi_2)$. Let π denote the mixing probability parameter of the mixture model and let z_i denote a latent mixture indicator variable such that $z_i = 1$ if y_i belongs to the first mixture component and $z_i = 0$ if it belongs to the second mixture component, for $i = 1, \ldots, n$. When the exact number of observations within each mixture component is fixed, the sum of z_i's is known as

$$\sum_{i=1}^{n} z_i = m, \quad 0 < m < n. \quad (2.1)$$

To deal with the information that the mixture component size is known, one may simply fix the value of the mixing probability parameter at $\pi = m/n$ and derive the EM algorithm accordingly; we call this a conventional EM algorithm throughout the paper. Then, the complete-data likelihood function for $\boldsymbol{\theta}$ is written as

$$L(\boldsymbol{\theta};\mathbf{y},\mathbf{z}) = \prod_{i=1}^{n} \left\{ f(y_i|\psi_1,\xi)^{z_i} f(y_i|\psi_2,\xi)^{1-z_i} \frac{m^{z_i}(n-m)^{1-z_i}}{n} \right\}.$$

Using the current iterate of parameters, $\boldsymbol{\theta}^{(t)} = \left(\xi^{(t)}, \psi_1^{(t)}, \psi_2^{(t)}\right)$, at iteration t, the E-step computes the conditional expectation of z_i given $\mathbf{y}$ and $\boldsymbol{\theta}^{(t)}$,

$$E\left(z_i|\mathbf{y},\boldsymbol{\theta}^{(t)}\right) = \frac{m f\left(y_i|\psi_1^{(t)},\xi^{(t)}\right)}{m f\left(y_i|\psi_1^{(t)},\xi^{(t)}\right) + (n-m) f\left(y_i|\psi_2^{(t)},\xi^{(t)}\right)} = \hat{z}_i^{(t+1)}.$$

Then the M-step sets

$$\boldsymbol{\theta}^{(t+1)} = arg \quad max_{\theta} \quad L\left(\boldsymbol{\theta};\mathbf{y},\hat{\mathbf{z}}^{(\mathbf{t+1})}\right),$$

where $\hat{\mathbf{z}}^{(t+1)} = \left(\hat{z}_1^{(t+1)}, \ldots, \hat{z}_n^{(t+1)}\right)$. The E-step and M-step are iterated until certain convergence criteria are satisfied.

2.2 The Proposed EM Algorithm

Despite the simplicity of its construction, the conventional EM algorithm does not deal with a correct model because the sum constraint in (2.1) is not fully accounted for. To improve the conventional EM algorithm by fully capitalizing on the available information, let us introduce a new latent vector $\mathbf{z}^* = (z_1^*, \ldots, z_n^*) \stackrel{d}{=}$ with support given by the set of all possible binary vectors on space

$$\mathscr{Z}^n = \left\{ (z_1, \ldots, z_n) : z_i = 0 \quad \text{or} \quad 1, \quad \text{and} \quad \sum_{i=1}^{n} z_i = m \right\}.$$

Then the random vector **z*** follows a so-called conditional Bernoulli distribution (Chen and Liu, 1997), and we write **z*** *CBe*(*n*, *m*, **p**) with probability mass function

$$p(\mathbf{z}^*) = \frac{\prod_{i=1}^{n} w_i^{z_i^*}}{R(m,S,\mathbf{w})},$$

where $w_i = p_i/(1 - p_i)$ is the odds of $z_i = 1$ with $p_i = P(z_i = 1)$, $S = \{1, \ldots, n\}$ is a set of indices, and $R(m,S,\mathbf{w}) = \Sigma_{B \subset S, |B|=m}(\Pi_{i \in B}\, w_i)$ is the normalizing constant with $R(0, S, \mathbf{w}) = 1$,

$R(m, S, \mathbf{w}) = 0$ for any $m > |S|$, and $\mathbf{w} = (w_1, \ldots, w_n)$. Then the complete-data likelihood function for $\boldsymbol{\theta}$ is written as

$$L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{z}^*) = \prod_{i=1}^{n} \left\{ f(y_i|\psi_1, \xi)^{z_i^*} f(y_i|\psi_2, \xi)^{1-z_i^*} \right\} p(\mathbf{z}^*), \quad (2.2)$$

where the probability of $z_i = 1$ is a priori set to $p_i = m/n$ for $i = 1, \ldots, n$.

Because the log of the complete-data likelihood function in (2.2) is linear in z_i^* with respect to $\boldsymbol{\theta}$, the E-step of the proposed EM algorithm amounts to computing the conditional expectation of z_i^* given $\mathbf{y}$ and $\boldsymbol{\theta}^{(t)}$. To derive the E-step, we compute the joint probability mass function of $\mathbf{z}^*$ given $\mathbf{y}$ and $\boldsymbol{\theta}^{(t)}$ as

$$\begin{aligned} p\left(\mathbf{z}^*|\mathbf{y}, \boldsymbol{\theta}^{(t)}\right) &= \frac{p\left(\mathbf{y}|\mathbf{z}^*, \boldsymbol{\theta}^{(t)}\right) p(\mathbf{z}^*)}{\sum_{\mathbf{z}^* \in \mathcal{Z}^n} p\left(\mathbf{y}|\mathbf{z}^*, \boldsymbol{\theta}^{(t)}\right) p(\mathbf{z}^*)} \\ &= \frac{\prod_{i=1}^{n} \left(w_i^{(t)}\right)^{z_i^*}}{R\left(m, S, \mathbf{w}^{(t)}\right)}, \end{aligned}$$

where $w_i^{(t)} = p_i^{(t)} / \left(1 - p_i^{(t)}\right)$ is the odds of $z_i^* = 1$ given $\mathbf{y}$ and $\boldsymbol{\theta}^{(t)}$, and

$$p_i^{(t)} = \frac{m f\left(y_i|\psi_1^{(t)}, \xi^{(t)}\right)}{m f\left(y_i|\psi_1^{(t)}, \xi^{(t)}\right) + (n-m) f\left(y_i|\psi_2^{(t)}, \xi^{(t)}\right)}.$$

Then the E-step of the proposed EM algorithm is given by

$$\begin{aligned} E\left(z_i^*|\mathbf{y}, \boldsymbol{\theta}^{(t)}\right) &= \frac{P\left(z_i^* = 1, \mathbf{z}_{-i}^*|\mathbf{y}, \boldsymbol{\theta}^{(t)}\right)}{P\left(\mathbf{z}_{-i}^*|z_i^* = 1, \mathbf{y}, \boldsymbol{\theta}^{(t)}\right)} \\ &= \frac{w_i^{(t)} R\left(m-1, S\backslash\{i\}, \mathbf{w}_{-i}^{(t)}\right)}{R\left(m, S, \mathbf{w}^{(t)}\right)} = \tilde{z}_i^{(t+1)}, \end{aligned} \quad (2.3)$$

where $\mathbf{z}_{-1}^* = \left(z_1^*, \ldots, z_{i-1}^*, z_{i+1}^*, \ldots, z_n^*\right)$ and $\mathbf{w}_{-1}^{(t)} = \left(w_1^{(t)}, \ldots, w_{i-1}^{(t)}, w_{i+1}^{(t)}, \ldots, w_n^{(t)}\right)$. The M-step of the proposed EM algorithm sets

$$\boldsymbol{\theta}^{(t+1)} = arg \quad max_\theta \quad L\left(\boldsymbol{\theta}; \mathbf{y}, \tilde{\mathbf{z}}^{(\mathbf{t+1})}\right),$$

where $\tilde{\mathbf{z}}^{(t+1)} = \left(\tilde{z}_1^{(t+1)}, \ldots, \tilde{z}_n^{(t+1)}\right)$.

3 Efficient and Numerically Stable Computation of the Proposed EM Algorithm

The $R(m, S, \mathbf{w})$ function is the summation over the product of all $\binom{n}{m}$ combinations of w_i's and the computation of the R function may not be practical because n is typically large. As proposed by Gail et al. (1981), we thus consider an efficient recursive method to calculate the summation. That is, for $S = \{1, \ldots, n\}$ and $1 \leq m \leq |S|$, we have

$$R(m, S, \mathbf{w}) = R(m, S \setminus \{i\}, \mathbf{w}_{-i}) + w_i R(m-1, S \setminus \{i\}, \mathbf{w}_{-i}). \quad (3.4)$$

Then $R(m, S, \mathbf{w})$ can be computed with $m(n-1)$ additions and $m(n-m+1)$ multiplications, which requires much less operations than $\binom{n}{m}$ evaluations.

In the context of fitting finite mixture models with known mixture component size, the computation of the R function can be numerically unstable in certain circumstances. First, when there exists a little overlap between the distributions of the mixture components, the probability of belonging to the first component given $\mathbf{y}$ and $\boldsymbol{\theta}^{(t)}$ can be close to one, so that the corresponding $w_i^{(t)}$ becomes extremely large. Because the R function in (2.3) is a sum of a product of $w_i^{(t)}$'s, such a large $w_i^{(t)}$ causes inflation of the R function and its computation can be numerically unstable. Second, when the sample size n is large, it is likely that some observations come from the tail of a distribution such that $p_i^{(t)}$ close to one and the corresponding $w_i^{(t)}$ becomes extremely large. Even when there are no such extreme observations, a product of relatively large $w_i^{(t)}$'s can still cause inflation of the R function, thereby making its computation numerically unstable.

To circumvent such numerical instability, we propose to cancel out a large common factor between the numerator and denominator in (2.3) to make its computation numerically stable by noting that the E-step is computed as the ratio of two R functions. Specifically, we factor out a product of some largest $w_i^{(t)}$'s and model the remaining expression of the R function. The modified R function, denoted by $\widetilde{R}(m, S, \mathbf{w})$, is defined as

$$\widetilde{R}(m, S, \mathbf{w}) = \frac{R(m, S, \mathbf{w})}{w_{[n-m+1]} w_{[n-m+2]} \cdots w_{[n]}},$$

where $w_{[1]} < w_{[2]} < \cdots < w_{[n]}$ denote the n order statistics based on $w_1, \ldots, w_n$, i.e., $\widetilde{R}(m, S, \mathbf{w})$ is the original $R(m, S, \mathbf{w})$ function divided by a product of the m largest w_i's.

Table 1 displays the arithmetic operations of $\widetilde{R}(m, S, \mathbf{w})$ for a simple example when $m = 2$, $S = \{1, 2, 3, 4\}$, and $\mathbf{w} = (w_1, w_2, w_3, w_4)$ with $w_4 < w_3 < w_2 < w_1$, i.e., $w_{[1]} = w_4$, $w_{[2]} = w_3$, $w_{[3]} = w_2$, and $w_{[4]} = w_1$. Starting from the upper-left corner of the table, $\widetilde{R}(m, S, \mathbf{w})$ is

generated at the lower-right corner. For $S = \{1, \ldots, n\}$, $1 \le m \le |S|$, and $1 \le j \le n$, the new efficient and numerically stable recursive method is given by

$$\widetilde{R}(m, S, \mathbf{w}) = \widetilde{R}(m, S \setminus \{i\}, \mathbf{w}_{-i}) + \widetilde{R}(m-1, S \setminus \{i\}, \mathbf{w}_{-i}) \frac{w_i}{w_{[n-j+1]}},$$

which requires the same number of operations as with the original efficient recursive method in (3.4) and thus the cost of computation remains the same, while greatly improving numerical stability. Using the new efficient recursive method, the E-step in (2.3) is rewritten as

$$E\left(z_i^* | \mathbf{y}, \boldsymbol{\theta}^{(t)}\right) = \begin{cases} \dfrac{w_i^{(t)} \widetilde{R}\left(m-1, S\setminus\{i\}, \mathbf{w}_{-i}^{(t)}\right)}{w_{[n-m+1]}^{(t)} \widetilde{R}\left(m, S, \mathbf{w}^{(t)}\right)} & \text{if} \quad w_i^{(t)} < w_{[n-m+1]}^{(t)} \\ \dfrac{\widetilde{R}\left(m-1, S\setminus\{i\}, \mathbf{w}_{-i}^{(t)}\right)}{\widetilde{R}\left(m, S, \mathbf{w}^{(t)}\right)} & \text{if} \quad w_i^{(t)} \ge w_{[n-m+1]}^{(t)}, \end{cases}$$

which is numerically stable for any given probability vector $\mathbf{p}^{(t)} = \left(p_1^{(t)}, \ldots, p_n^{(t)}\right)$.

4 Comparison of the Conventional and Proposed EM Algorithms

We want to make clear the issues involved in estimation for our proposed EM algorithm in comparison to the conventional EM algorithm when we know the exact numbers of members in each component. To do this, we examine in detail the case where observations are taken from a mixture of two normal distributions, $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$. In the three parameter setting (μ_1, μ_2, σ) of the two-component normal mixture model with known π, the maximum likelihood estimate exists and is consistent (Basford and McLachlan, 1985; McLachlan and Peel, 2000). However, Nettleton (1999) notes that the estimate of the conventional EM algorithm in this context (for example by Gould and Shih (1992) who use $\pi = m/n$) may converge toward the boundary of the parameter space while a true parameter lies in the interior of the parameter space. Friede and Kieser (2002) also show through a simulation study that the estimate of a within-group standard deviation from the conventional EM algorithm depends on the starting value of a standardized treatment effect, $\delta = (\mu_2 - \mu_1)/\sigma$. Here we further illustrate the effect of the starting values on the convergence behavior of both the conventional and proposed EM algorithms, and the reason why the conventional EM algorithm sometimes converges toward the boundary of the parameter space.

For a simulation study, we generate 1000 test data sets, each of size 20, where 10 observations are from $y_i \sim N(\mu_1, \sigma)$ for $i = 1, \ldots, 10$, the other 10 observations are from $y_i \sim N(\mu_2, \sigma)$ for $i = 11, \ldots, 20$, and the true values of parameters are set to $(\mu_1, \mu_2, \sigma) = (0, 1, 1)$. Both the conventional and proposed EM algorithms are used to estimate (μ_1, μ_2, σ) with identical stopping rules. To demonstrate the dependence of convergence on starting values, we randomly select one out of the 1000 test data sets and run both the conventional and proposed EM algorithms with 40 different starting values

$\left(\mu_1^{(0)}, \mu_2^{(0)}, \sigma^{(0)}\right) = (\overline{y} - d, \overline{y} + d, 4)$, where d = 0.005, 0.105, …, 3.905. Figure 1 shows the EM estimates for (μ_1, μ_2, σ) as a function of half the starting value $\delta^{(0)} = \left(\mu_2^{(0)} - \mu_1^{(0)}\right)/\sigma^{(0)}$. As shown in Figure 1, the conventional EM algorithm converges toward the boundary of the parameter space when $\delta^{(0)}$ is small and to the interior mode when $\delta^{(0)}$ is sufficiently large. That is, the conventional EM estimate depends on the starting value $\delta^{(0)}$. From the other test data sets, we also note that a cutoff value for $\delta^{(0)}$ that makes the conventional EM algorithm converge to the interior mode varies from data to data. By contrast, the proposed EM algorithm always converges to the interior mode, depending on neither $\delta^{(0)}$ nor the type of data.

Figure 2 shows the scatterplot of 1000 EM estimates of (μ_1, μ_2) for the conventional and proposed EM algorithms with identical starting values $\delta^{(0)}$ = 3. As shown in the left panel of Figure 2, the conventional EM estimates compose two apparent clusters with one near the true value (0, 1) and the other toward the boundary of the parameter space. The boundary of the parameter space implies that the underlying distribution is from a single component, which is incorrect because there exist two mixture components with separate means. Such artificial modes near the boundary of the parameter space may occur because the conventional EM algorithm deals with an incorrect complete-data likelihood without fully accounting for the fact that the exact number of observations within each mixture component is fixed. By contrast, the proposed EM estimates are all nicely spread out around the true value (0, 1).

We empirically explore the reason why the conventional EM algorithm sometimes produces boundary estimates for (μ_1, μ_2), while the proposed EM algorithm does not. To illustrate the reason, we used the same data set as in Figure 1. In the M-step of both EM algorithms, the means and standard deviation are conditionally maximized, i.e., we iterate between the maximization of (μ_1, μ_2) given σ and the maximization of σ given (μ_1, μ_2). Thus we examine the profile log-likelihood function of the conventional and proposed EM algorithms. Figure 3 shows the behavior of the profile log-likelihood functions of the two EM algorithms when they start with an extreme starting value that is on the boundary of the parameter space. When it comes to the conventional EM algorithm, Figure 3(a) shows the profile log-likelihood function for σ given $(\mu_1^{(0)}, \mu_2^{(0)})$ such that the starting values for the means are given by $\mu_1^{(0)} = \mu_2^{(0)} = \Sigma_{i=1}^{20} y_i/20$. Note that the starting value of $\mu_1^{(0)} = \mu_2^{(0)}$ is rather extreme because it implies that the underlying distribution is from a simple component; we use the extreme starting value for illustration purposes. The profile log-likelihood function is maximized at $\sigma^{(1)}$ = 1.33 given $\left(\mu_1^{(0)}, \mu_2^{(0)}\right)$. Figure 3(b) shows the profile log-likelihood function for (μ_1, μ_2) given $\sigma^{(1)}$ = 1.33, which has a single mode at $\mu_1 = \mu_2$. Thus the profile log-likelihood function for ($\mu_1 = \mu_2$) is necessarily maximized $\mu_1^{(1)} = \mu_2^{(1)}$, which implies that $\sigma^{(2)}$ that maximizes the profile log-likelihood function for σ given $\left(\mu_1^{(1)} = \mu_2^{(1)}\right)$ remains at $\sigma^{(1)}$. Because of being trapped near the boundary of the parameter space, the conventional EM algorithm does not find the interior mode when it begins with $\mu_1^{(0)} = \mu_2^{(0)}$ in this particular data set.

In the case of the proposed EM algorithm, Figure 3(c) shows the profile log-likelihood function for σ given $\left(\mu_1^{(0)}, \mu_2^{(0)}\right)$ such that the starting values for the means are also given by $\mu_1^{(0)}=\mu_2^{(0)}=\Sigma_{i=1}^{20} y_i/20$. The profile log-likelihood function for σ is then maximized at $\sigma^{(1)}$ = 1.33, but the profile log-likelihood function for (μ_1, μ_2) given $\sigma^{(1)}$ = 1.33 now becomes slightly bimodal, as shown in Figure 3(d). Within the parameter space for (μ_1, μ_2), the profile log-likelihood function for ($\mu_1 = \mu_2$) given σ = 1.33 is maximized at $\left(\mu_1^{(1)}, \mu_2^{(1)}\right) = (0.35, 0.9)$. Next, Figure 3(e) shows that the profile log-likelihood function for σ given $\left(\mu_1^{(1)}, \mu_2^{(1)}\right) = (0.35, 0.9)$ is maximized at $\sigma^{(2)}$ = 1.29. Then the profile log-likelihood surface for (μ_1, μ_2) given $\sigma^{(2)}$ = 1.29 has two modes which are further apart, as shown in Figure 3(f). By alternating the conditional maximization steps, the estimates of (μ_1, μ_2) move away from the boundary near $\mu_1 = \mu_2$. This example illustrates how the proposed EM algorithm obtains the interior mode when the conventional EM algorithm cannot.

5 An Adaptive Design Clinical Trial Example

We illustrate the proposed EM algorithm and compare it with the conventional EM algorithm by applying them to the design of a realistic sample-size adaptive clinical trial. To evaluate the properties of an adaptive design in actual practice, we use a simulation study for a randomized trial of an experimental compound versus an appropriate comparator. The primary endpoints are assumed to follow two different normal distributions with a common standard deviation σ. Suppose we want a power of 80% to detect a clinically meaningful treatment difference with a level 0.05 test, assuming the endpoint has a standard deviation σ = 1 in each treatment group. This requires an initial sample size of N = 160, i.e., a total of 160 observations in the trial. We plan to apply block randomization to avoid imbalance across the trial. In general practice, block sizes are randomly chosen (e.g., as a sequence of 4, 6, 2, 4, etc) to protect against possibly guessing the next patient's treatment allocation when small block sizes are used. When planning our trial, we chose a constant block size of 4 as a simplified scenario for illustrative purposes.

Our adaptive clinical trial plans an interim analysis based on the first n = 80 completed patients with 20 randomization blocks, where m = 40 patients are assigned to the experimental group. Both the conventional and proposed EM algorithms are used to estimate σ from these n observations without knowing the treatment identities, and then recalculate the total sample size N' at the interim. Then the additional $N' - n$ observations are simulated to generate a total of N' observations in the trial. We repeat this procedure 3000 times to examine the power based on a sample of size N' calculated by each EM algorithm's interim estimate of σ. The expected size of additional samples is calculated by the average of the simulated additional sample sizes.

The initial sample size N needs to be increased if σ is incorrectly assumed too low; see Teel (2011) for details. We thus assume that σ was initially underestimated as σ = 1 when the true value of σ is $\sqrt{2}$. Then we examine how the blinded sample size re-estimation mitigates the effect of a false assumption of σ on the power of a trial. When an initial sample size is calculated as N = 160 with an underestimated value of σ = 1, the fixed sample size design

can only achieve a power of 51% at the true value of $\sigma = \sqrt{2}$. Our power simulation study shows that by using the conventional EM procedure, the average number of observations required to attain a power of 80% is computed as $N' = 248$ and the corresponding average power is 65.6% at the true value of $\sigma = \sqrt{2}$. By contrast, the proposed EM procedure yields an adjusted sample size of $N' = 298$ on average and the corresponding average power is 73% at the true value of $\sigma = \sqrt{2}$. Although σ is underestimated as two-third the true value in the planning phase of the study, we obtain a power of 73% using the sample size adjustment based on the proposed EM procedure with block size 4. Furthermore, we note that the expected size of adjusted total samples, $N' = 298$, is close to 320 which is the fixed sample size required to achieve a power of 80% at the true value of $\sigma = \sqrt{2}$. Interestingly, the fixed sample power at 298 patients is 76.7%, so that the "cost" of the adaptive design is a "loss" of less that 4% in power compared to a fixed sample size design that knows the true value of σ. Since the goal is to maintain power at 80%, the proposed EM procedure is considered to be superior to the conventional EM procedure. These results are typical of a variety of scenarios (Teel, 2011) and demonstrate that our proposed EM algorithm is more advantageous than the conventional EM algorithm in maintaining power when doing sample size re-estimation.

6 Discussion

In this paper, we propose a new EM algorithm for two-component mixture models when the exact numbers of observations in each mixture component are given. The E-step of the new EM algorithm involves a conditional Bernoulli distribution, which requires the computation of the normalizing constant of a conditional Bernoulli distribution, denoted by $R(m, S, \boldsymbol{w})$. Because a naive way of calculating $R(m, S, \boldsymbol{w})$ is computationally expensive, we consider using the recursive method for calculating the normalizing constant of a conditional Bernoulli distribution, proposed by Gail et al. (1981). However, the recursive method used by Gail et al. (1981) can be numerically unstable in the context of mixture modeling. Thus we propose a new recursive method for calculating the normalizing constant of a conditional Bernoulli distribution that is not only efficient but also numerical stable. By fully accounting for the fact that the number of observations within each mixture component is given, the proposed EM algorithm produces maximum likelihood estimates that are robust to starting values and correctly lie in the interior of the parameter space.

When developing the new EM algorithm for mixture models with known mixture component size, the number of mixture components is confined to two. Although it is more general and practical to assume more than two components in the mixture model, the method for the general finite mixture models can be derived by a nontrivial generalization of the method proposed in this paper. In an adaptive randomized clinical trial that is an important application of our proposed method, it is also not uncommon to compare only the treatment and control groups. Thus, our paper focuses on the two-component mixture models and methods for the general finite mixture models are to be dealt with in future research.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation grant (KOSEF-2012-8-0575) funded by the Korea government and the NIMH grant (1-P50-MH084053).

References

Basford KE, McLachlan GJ. Likelihood estimation with normal mixture models. Journal of Applied Statistics. 1985; 34:282–289.

Chen X, Dempster AP, Liu J. Weighted finite population sampling to maximize entropy. Biometrika. 1994; 81:457–469.

Chen X, Liu J. Statistical application of the poisson-binomial and conditional Bernoulli distributions. Statistica Sinica. 1997; 7:875–892.

Chen Y, Yang S. Estimating disaggregate models using aggregate data through augmentation of individual choice. Journal of Marketing Research. 2007; 44:613–621.

Choi T, Schervish MJ, Schmitt KA, Small MJ. A Bayesian approach to logistic regression model with incomplete information. Biometrics. 2008; 64:424–430. [PubMed: 17764482]

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

Friede T, Kieser M. A comparison of methods for adaptive sample size adjustment. Statistics in Medicine. 2001; 20:3861–3873. [PubMed: 11782039]

Friede T, Kieser M. On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. Statistics in Medicine. 2002; 21:165–176. [PubMed: 11782057]

Gail MH, Lubin JH, Rubinstein LV. Likelihood calculation for matched case-control studies and survival studies with tied death times. Biometrika. 1981; 68:703–707.

Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Communications in Statistics. 1992; 21:2833–2853.

McLachlan, GJ.; Peel, D. Finite mixture models. Wiley; New York: 2000.

Musalem A, Bradlow ET, Raju JS. Bayesian estimation of random-coefficients choice models using aggregate data. Journal of Applied Econometrics. 2009; 24:490–516.

Nettleton D. Convergence properties of the EM algorithm in constrained parameter spaces. The Canadian Journal of Statistics. 1999; 27:639–648.

Park T. Bayesian analysis of individual choice behavior with aggregate data. The Journal of Computational and Graphical Statistics. 2011; 20:158–173.

Rosenberger, WF.; Lachin, JM. Randomization in clinical trials: theory and practice. Wiley; New York: 2002. p. 81-83.

Stephens M. Dealing with label switching in mixture models. Journal of Royal Statistical Society. 2000; 62:795–809.

Teel, C. Ph.D. thesis. Department of Statistics, University of Pittsburgh; 2011. Improved sample size reestimation in adaptive clinical trials without unblinding.

Verhelst ND. An efficient MCMC algorithm to sampler binary matrices with fixed marginals. Psychometrika. 2008; 73:705–728.

Xing B, Ganju J. A method to estimate the variance of an endpoint from an on-going blinded trial. Statistics in Medicine. 2005; 24:1807–1814. [PubMed: 15803440]

Zucker DM, Wittes JT, Schabenberger O, Brittain E. Internal pilot studies II: comparison of various procedure. Statistics in Medicine. 1999; 18:3493–3509. [PubMed: 10611621]

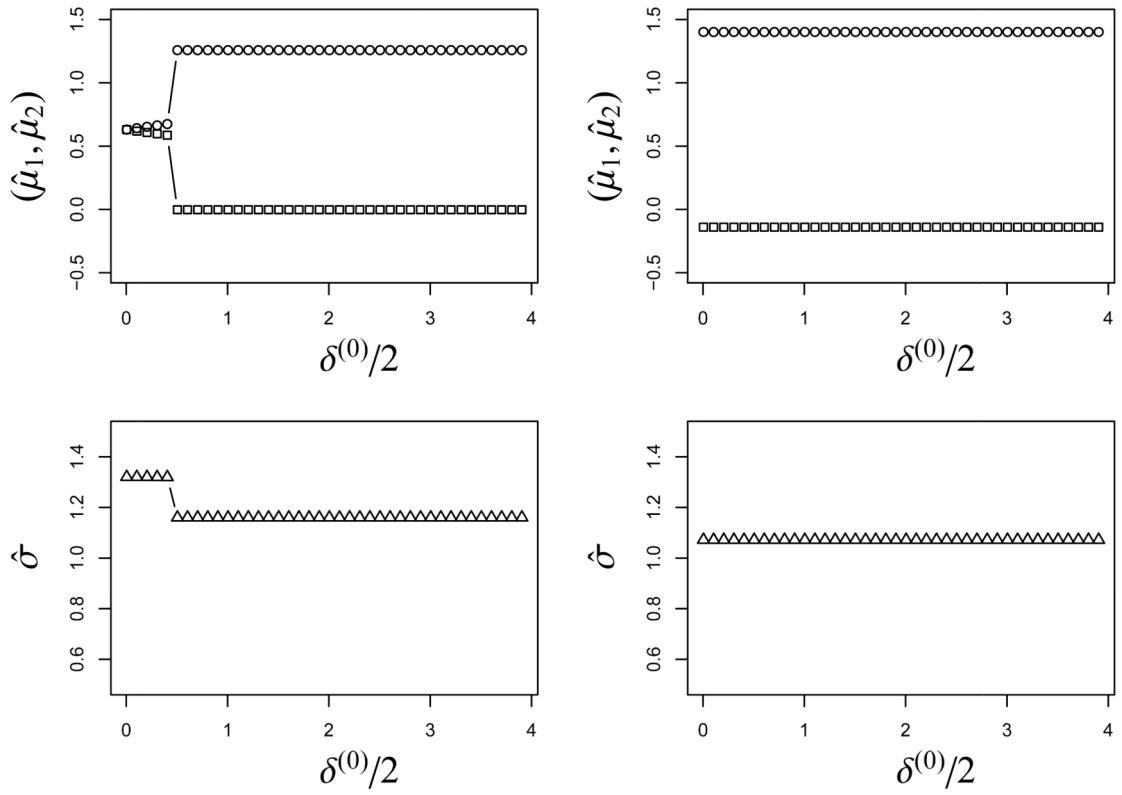

Figure 1.

EM estimates for parameters, $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})$, as a function of half the starting value of a standardized treatment effect. The left panels correspond to the conventional EM algorithm, while the right panels correspond to the proposed EM algorithm.

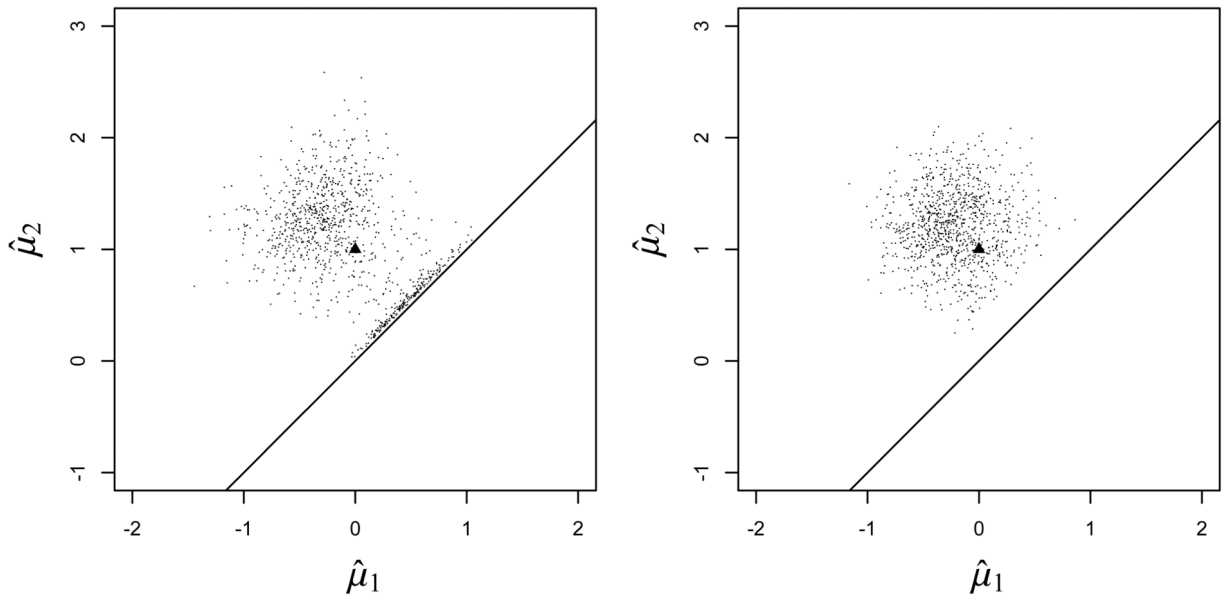

Figure 2.

Comparison of two EM estimates for μ_1 and μ_2 using 1000 test data sets. The triangle indicates the true value of $(\mu_1, \mu_2) = (0, 1)$. The left panel is the result from the conventional EM algorithm and the right panel is from the proposed EM algorithm. The parameter space lies above the 45-degree line.

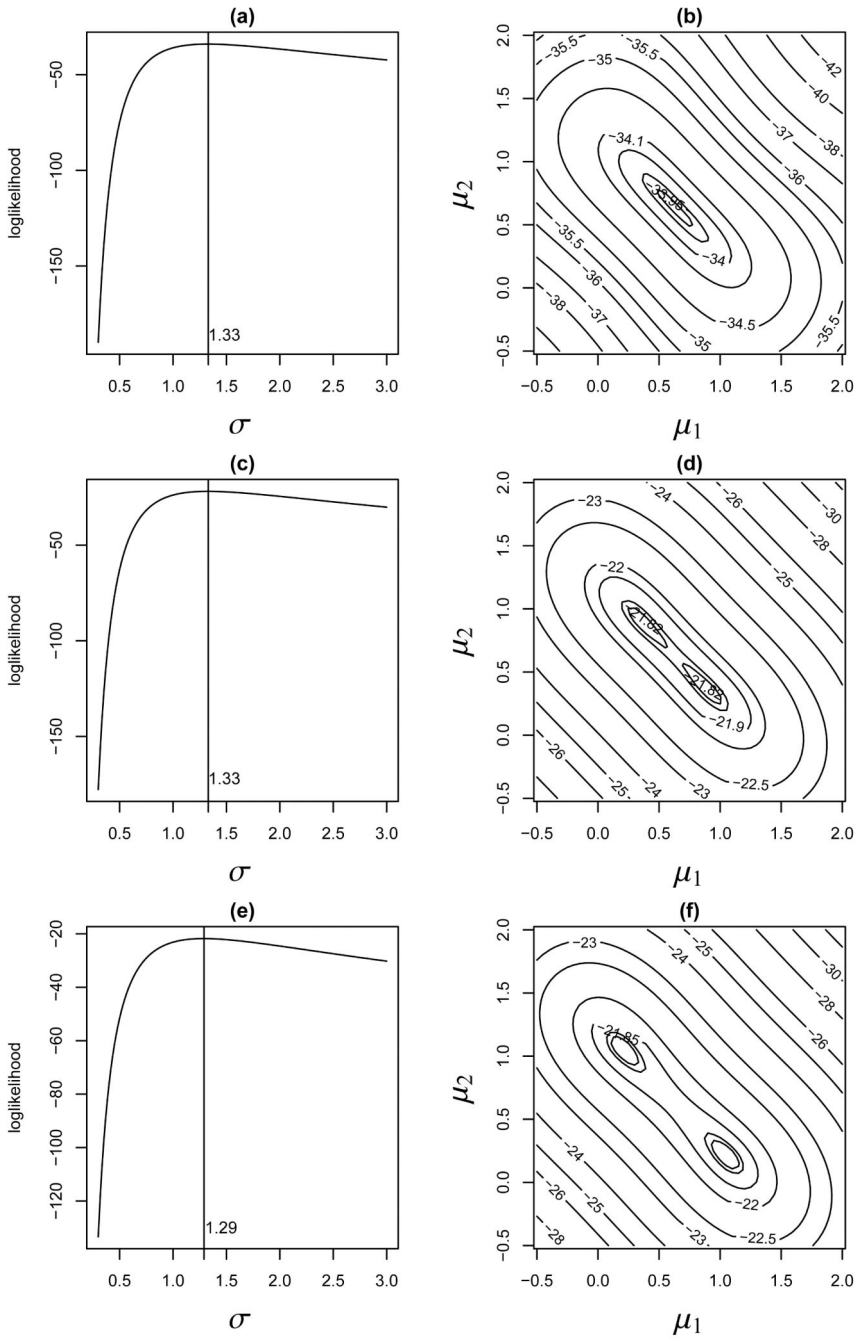

Figure 3.
Profile log-likelihood functions of the conventional and proposed EM algorithms.

Table 1

New recursive generation of $\widetilde{R}(m, S, \mathbf{w})$ when $m = 2$, $S = \{1, 2, 3, 4\}$, and $\mathbf{w} = (w_1, w_2, w_3, w_4)$ with $w_4 < w_3 < w_2 < w_1$.