ELSEVIER Computational Statistics & Data Analysis 41 (2003) 591–601

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Asymptotic properties of the EM algorithm estimate for normal mixture models with component specific variances

Dechavudh Nityasuddhi[a,*], Dankmar Böhning[b]

[a] *Department of Biostatistics, Faculty of Public Health, Mahidol University, 420/1 Rajvithee, Rachathewi, Bangkok 10400, Thailand*

[b] *Department of Epidemiology, Free University Berlin, Fabeckstr. 60-62, 14195 Berlin, Germany*

Received 1 February 2002; received in revised form 1 March 2002

Abstract

Most of the researchers in the application areas usually use the EM algorithm to find estimators of the normal mixture distribution with unknown component specific variances without knowing much about the properties of the estimators. It is unclear for which situations the EM algorithm provides "good" estimators, good in the sense of statistical properties like consistency, bias, or mean square error. A simulation study is designed to investigate this problem. The scope of this study is set for the mixture model of normal distributions with component specific variance, while the number of components is fixed. The asymptotic properties of the EM algorithm estimate is investigated in each situation. The results show that the EM algorithm estimate does provide good asymptotic properties except for some situations in which the population means are quite close to each other and larger differences in the variances of the component distributions occur.

Keywords: Normal mixture model; Component specific variances; EM algorithm; Simulation study

1. Introduction

In many applications in medicine, biology, or the social sciences situations are studied in which data such as systolic blood pressure, blood cholesterol level and glucose level are of the following type: under standard assumptions the population is homogeneous, leading to a simple, one or two parametric and natural density. Examples include the binomial, the Poisson, the geometric, the exponential and the normal

* Corresponding author.
E-mail address: phdnt@mahidol.ac.th (D. Nityasuddhi).

0167-9473/03/$ - see front matter
PII: S0167-9473(02)00176-7

distribution. We consider a most common one, the *normal distribution*. If these standard assumptions are violated because of population heterogeneity, mixture models can capture these additional complexities easily. The mixture model arises as the marginal distribution over the (latent) variable describing sub-population membership. In this setting here, we assume that the number of sub-populations g is known, leading to a discrete non-parametric mixing distribution. Moreover, we allow the mixture kernel to be normal with *component-specific variance*. Maximum-likelihood estimation with the EM algorithm has been the approach to the mixture density estimation problem, which is a particular iterative procedure for numerically, iteratively approximating maximum-likelihood estimates of the parameters in mixture densities. This procedure is a specialization to the mixture density estimation problem of a general method for a approximating iteratively maximum-likelihood estimates in an incomplete data context which was formalized by Dempster et al. (1977) and termed by them the EM algorithm (E for "expectation" and M for "maximization"). It has been found in most instances to have the advantage of rather reliable convergence, low cost per iteration, economy of storage and ease of programming, as well as a certain heuristic appeal; unfortunately, its convergence can be slow even in simple problems (Redner and Walker, 1984; Meng, 1997).

Nevertheless, the practitioners in applied statistics and elsewhere make intensive use of the EM algorithm which is likely to provide some local maximum of the likelihood function and the estimator might be considered as some form of local maximum-likelihood estimator (Behboodian, 1970). It is, however, by no means guaranteed that the EM algorithm provides a global maximum (Wu, 1983). In addition, in the case of normal mixtures with component-specific variances, the log-likelihood is unbounded and attains $+\infty$ for certain values of the parameter space. Whereas algorithmic approaches of global character such as gradient function based techniques (Böhning, 2000) fail miserably in this case ("they climb up the hill for ever"), the *local* character of the EM algorithm adds to its advantage—as many practitioners feel that the EM algorithm provides rather reasonable solutions. Though used much, surprisingly little theoretical knowledge is available for this estimator. In fact, it might be unclear to which extent asymptotic properties of the estimator such as consistency, asymptotic efficiency and asymptotic normality hold.

As there are no finite parameter values existing which maximize the likelihood, in our opinion, it is only fair not to speak about maximum-likelihood estimates, but rather about the estimates which the EM algorithm provides (some sort of solution of the score equation) and call them *EM algorithm estimates*. The problem is quite well known in the literature. McLachlan and Peel (2000, p. 41) write in connection with this notational problem: "We shall henceforth refer to $\hat{\Psi}$ as the MLE even in situations where it may not globally maximize the likelihood. Indeed, in some of the examples on mixture models to be presented, the likelihood is unbounded. However, for these models there may still, under regularity conditions, a sequence of roots of the likelihood equation corresponding to local maxima with the properties of consistency, efficiency, and asymptotic normality."

Since the EM algorithm has to be started with certain values for the parameters it is optimizing, the answers will depend to a certain degree on the initial values used

to start it (Böhning, 2000). Therefore, emphasis will be also given to the strategy of choosing initial values.

In the following we briefly describe the EM algorithm, the initial value strategy, the design of the simulation study and the results in terms of Bias and MSE.

2. EM algorithm for normal mixture model

The EM algorithm is a general-purpose algorithm to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data. It consists of an expectation step followed by a maximization step (Dempster et al., 1977). We assume that $y_1,\ldots,y_n$ are independent random variables each having a discrete mixture of g normal distributions with mean parameter μ_i and variance parameter σ_i^2. Denote with $p_1,\ldots,p_g;\ \mu_1,\ldots,\mu_g;\ \sigma_1^2,\ldots,\sigma_g^2$ all the unknown parameters in these g normal component densities, and let Ψ contain all of the unknown parameters. Note that there are $3g-1$ independent parameters, since the weights $p_1,\ldots,p_g$ have to sum to 1. Then, the incomplete-data log-likelihood function for Ψ is given by

$$\log L_{\mathrm{I}}(\Psi)=\sum_{j=1}^{n}\log\left\{\sum_{i=1}^{g}p_i f_i(y_j;\mu_i,\sigma_i^2)\right\},\quad \text{where } f_i(y_j;\mu_i,\sigma_i^2)$$

denotes the normal distribution with mean μ_i and variance σ_i^2.

For the purpose of the application of EM algorithm, the observed-data $y_{\mathrm{obs}}=(y_1,\ldots,y_n)$ are regarded as being incomplete. The latent variables z_{ij} are introduced, where z_{ij} is defined to be one or zero according to whether y_j did or did not arise from the ith component of mixture model ($i=1,\ldots,g;\ j=1,\ldots,n$). So, the complete-data x_{c} is given by $x_{\mathrm{c}}=(x_1,\ldots,x_n)$, where $x_1=(y_1,z_1),\ldots,\ x_n=(y_n,z_n)$ are taken to be independent and identically distributed with $z_1,\ldots,z_n$ being independent from a multinomial distribution consisting of a draw on g categories with respective probabilities $p_1,\ldots,p_g$. For this specification, the complete-data log likelihood is

$$\log L_{\mathrm{c}}(\Psi)=\sum_{i=1}^{g}\sum_{j=1}^{n}z_{ij}\log\{p_i f_i(y_j;\mu_i,\sigma_i^2)\}.$$

The EM algorithm is easy to program and proceeds iteratively in two steps, E (for expectation) and M (for maximization) (McLachlan and Krishnan, 1997). On the $(k+1)$st iteration, the E-step requires the calculation of the conditional expectation of the complete-data log-likelihood $\log L_{\mathrm{c}}(\Psi)$, given the observed data y_{obs}, using current fit $\Psi^{(k)}$ for Ψ.

$$Q(\Psi;\Psi^{(k)})=E_{\Psi^{(k)}}\{\log L_{\mathrm{c}}(\Psi)\mid y_{\mathrm{obs}}\}.$$

Since $\log L_{\mathrm{c}}(\psi)$ is a linear function of the unobservable component label variables z_{ij}, the E-step is calculated simply by replacing z_{ij} by its conditional expectation given y_j, using $\psi^{(k)}$ for ψ. That is, z_{ij} is replaced by

$$\begin{aligned}z_{ij}^{(k)}&=E_{\Psi^{(k)}}\{Z_{ij}|y_j\}\\&=\frac{p_i^{(k)}f(y_j;\mu_i^{(k)},\sigma_i^{2(k)})}{\sum_{m=1}^{g}p_m^{(k)}f(y_j;\mu_m^{(k)},\sigma_m^{2(k)})},\end{aligned}$$

which is the estimate after the kth iteration of the posterior probability that the jth entity with y_j belongs to the ith component ($i = 1,\ldots,g;\ j = 1,\ldots,n$).

The M-step requires the computation of $\hat{p}_i^{(k+1)}$, $\hat{\mu}_i^{(k+1)}$, $\sigma_i^{2(k+1)}$ ($i = 1,2,\ldots,g$) by maximization $\log L_c(\Psi)$; it is equivalent to computing the sample proportion, the weighted sample mean and sample variance with weight z_{ij}.

As $\log L_c(\Psi)$ is linear in the z_{ij}, it follows that z_{ij} are replaced by their conditional expectations $z_{ij}^{(k)}$. On the $(k+1)$th iteration, the intent is to choose the value of Ψ, say $\Psi^{(k+1)}$, that maximizes $Q(\Psi;\Psi^{(k)})$. It follows that on the M-step of the $(k+1)$st iteration, the current fit for the mixing proportions, the component means, and the variances is given explicitly by

$$\hat{p}_i^{(k+1)} = \sum_{j=1}^{n} z_{ij}^{(k)}/n,$$

$$\hat{\mu}_i^{(k+1)} = \sum_{j=1}^{n} z_{ij}^{(k)} y_j \Bigg/ \sum_{j=1}^{n} z_{ij}^{(k)},$$

$$\hat{\sigma}_i^{2(k+1)} = \sum_{j=1}^{n} z_{ij}^{(k)} (y_j - \hat{\mu}_i^{(k)})^2/n$$

for $i=1,\ldots,g$. The E- and M-steps are alternated repeatedly until the likelihood change by some small amount (to be set) in the case of convergence.

3. Simulation

We define the measures for the evaluation of the asymptotic properties of the estimators Ψ as follows.

1. Bias of the estimate Ψ for the ith component

$$\mathrm{BIAS}(\hat{\Psi}_i) = \frac{1}{r}\sum_{m=1}^{r} \hat{\Psi}_i^{(m)} - \Psi_i,$$

where $\hat{\Psi}_i = [\hat{p}_i, \hat{\mu}_i, \hat{\sigma}_i^2]$, $\Psi_i = [p_i, \mu_i, \sigma_i^2]$, and r is the number of simulation runs $\hat{\Psi}_i^{(m)} = [\hat{p}_i^{(m)}, \hat{\mu}_i^{(m)}, \hat{\sigma}_i^{2(m)}]$ of the mth iteration with $m = 1,\ldots,r$.

2. Mean square error (MSE) of the estimate Ψ_i for the ith component

$$\mathrm{MSE}(\hat{\Psi}_i) = \frac{1}{r}\sum_{m=1}^{r} (\hat{\Psi}_i^{(m)} - \Psi_i)^2.$$

In this study, we use FORTRAN 90 with IMSL library to develop the simulation program. We simulate 2–5 components of normal data with the combination of mean μ (equal to $5, 10, 20, 50$, and 100), variance σ^2 (equal to $1, 2, 5, 10, 20, 50$, and 100) and the mixing weight p of 0.1(0.1)0.9. Since there are a large number of combination values of the three parameters, we combine these into one index D to have a measure in the heterogeneity of these parameters.

Table 1
Typical combinations of Ψ for small, medium and large value of D

Gp.	D	p_1	μ_1	σ_1^2	p_2	μ_2	σ_2^2	p_3	μ_3	σ_3^2	p_4	μ_4	σ_4^2	p_5	μ_5	σ_5^2
2	Small	0.5	5	10	0.5	20	20									
	Medium	0.5	50	50	0.5	100	10									
	Large	0.5	10	100	0.5	100	20									
3	Small	1/3	5	1	1/3	10	5	1/3	20	10						
	Medium	1/3	5	100	1/3	20	100	1/3	50	50						
	Large	1/3	5	20	1/3	10	20	1/3	100	100						
4	Small	0.25	10	5	0.25	10	20	0.25	20	10	0.25	20	20			
	Medium	0.25	20	20	0.25	20	20	0.25	50	20	0.25	100	20			
	Large	0.25	5	100	0.25	20	100	0.25	50	50	0.25	100	1			
5	Small	0.2	5	2	0.2	5	10	0.2	5	20	0.2	10	20	0.2	20	5
	Medium	0.2	20	20	0.2	50	50	0.2	20	50	0.2	50	20	0.2	50	100
	Large	0.2	10	10	0.2	20	10	0.2	50	100	0.2	100	50	0.2	100	100

We define for a given vector Ψ

$$D = \sum_{i=1}^{g} p_i[(\mu_i - \bar{\mu})^2 + (\sigma_i^2 - \bar{\sigma}^2)^2],$$

where $\bar{\mu} = p_1\mu_1 + \cdots + p_g\mu_g$ and $\bar{\sigma}^2 = p_1\sigma_1^2 + \cdots + p_g\sigma_g^2$. If μ_i's are equal, then $D = \sum_{i=1}^{g} p_i(\sigma_i^2 - \bar{\sigma}^2)^2$. If σ_i^2's are equal, then $D = \sum_{i=1}^{g} p_i(\mu_i - \bar{\mu})^2$. If μ_i's are equal as well as the σ_i^2's, then $D = 0$. To interpret D, note the following. If one looks at the bivariate distribution giving mass p_i to the g 2-vectors $(\mu_i, \sigma_i^2)^{\mathrm{T}}$, then D represents just the *trace* of the covariance matrix of this distribution. Other possible measures would be the *determinant*, though we are not proceeding in this direction.

For each simulation set we fixed the total size (n) of the all components from 25 to 10,000 as $n = 25$, 50, 100, 200, 500, 1000, 2000, 5000, and 10,000. Then do the following steps of simulation process and repeat it 1000 times.

Steps in each simulation process:

1. Create a data set of size n
   1.1. Use multinomial distribution to generate the size of each component.
   1.2. Generate normal data for each component
   1.3. Combine the normal data sets into one ordered data set, y_i of size n
2. Setting initial values
   2.1. Partition y_j into g components with size greater than one by slide the $g-1$ cut points for all possible partitions. For $g = 2, 3, 4$, and 5, we have the number of partition $t = n - 3$, $t = \sum_{j=5}^{n-1}(n-j)$, $t = \sum_{i=1}^{n-7}\sum_{j=i+6}^{n-1}(n-j)$, and $t = \sum_{k=1}^{n-9}\sum_{i=k}^{n-9}\sum_{j=i+8}^{n-1}(n-j)$, respectively.
   2.2. In each partition set, compute $\hat{p}_i, \hat{\mu}_i, \hat{\sigma}_i^2$ and use it as initial values.

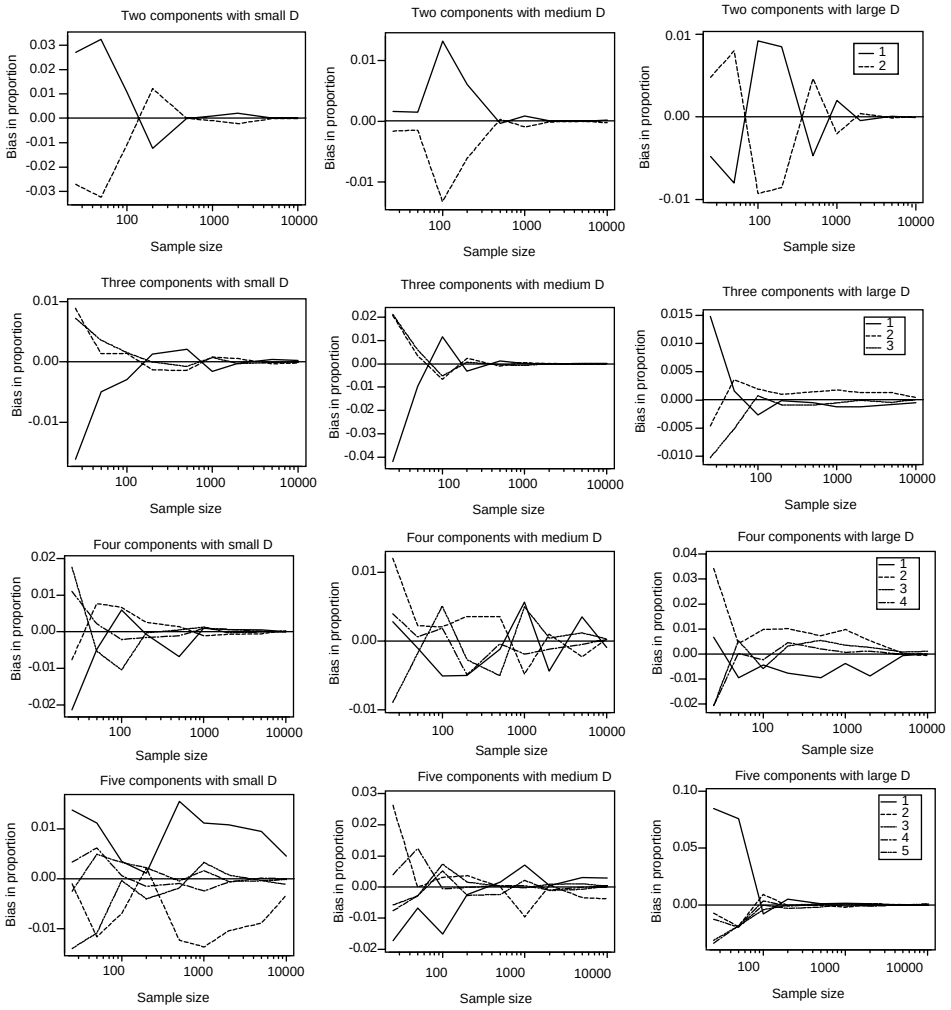

Fig. 1. Bias in proportion for two to five components with the small, medium and large value of D.

3. For each partition set do the EM-algorithm to find $L(\hat{\Psi}_i^{(q)})$, $\hat{\Psi}_i^{(q)}$ where $q = 1, 2, \ldots, t$ and select $\hat{\Psi}_i^{(m)} = \hat{\Psi}_i^{(q)}$ which gives the maximum value of $L(\hat{\Psi}_i^{(q)})$ for $q = 1, 2, \ldots t$.
4. Repeat step 1–3, 1000 times and compute BIAS($\hat{\Psi}_i$), MSE($\hat{\Psi}_i$), and D.

4. Results

As the index D combines mixing proportion, mean and variance of the normal data sets, we can classify D into small, medium and large value. The small value of D represents mixture of normal data that overlap largely, while the medium value of D

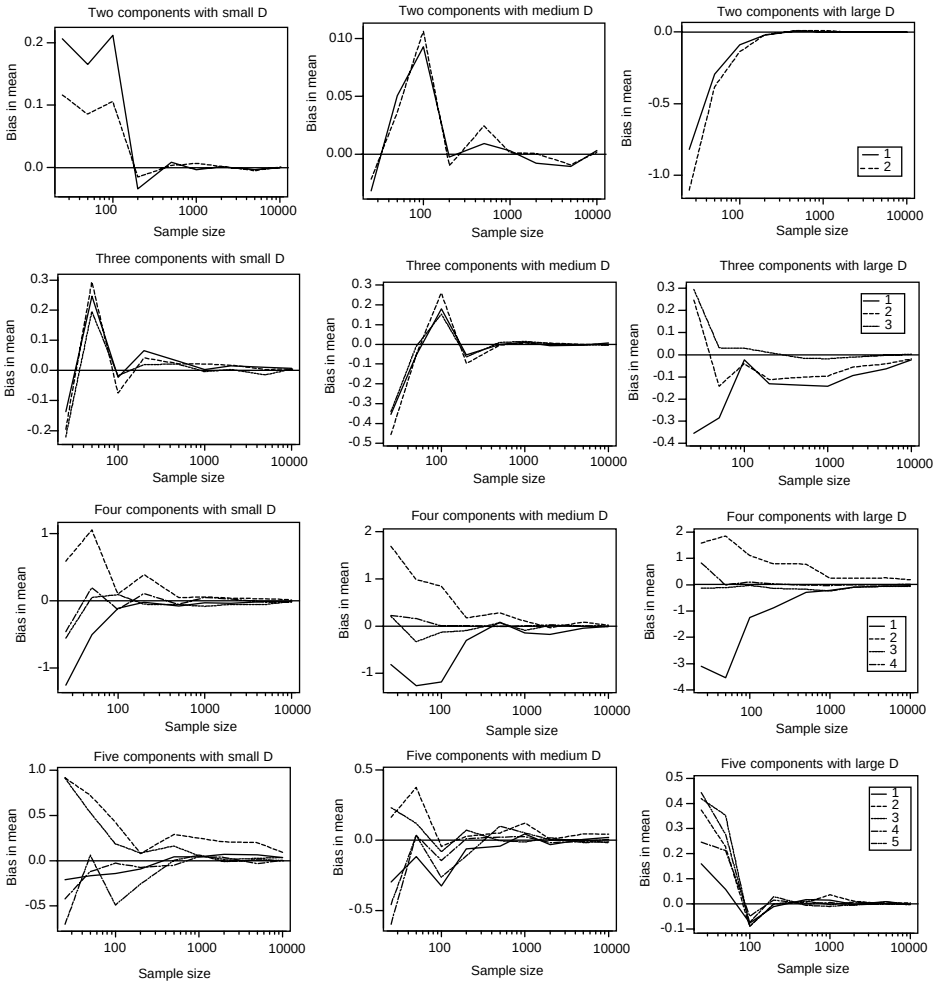

Fig. 2. Bias in mean for two to five components with small, medium and large value of D.

comes from mixtures of normal data that have some overlap. When the normal data sets show only slight overlap D has a large value. For this study we set D about 100 for the small, about 1000 for the medium and about 3000 for the large one. Some typical combinations for Ψ are shown in Table 1. After all, we run these simulations and provide the results in Figs. 1–6.

Consider the bias in the mean. Here, the estimates with medium and large D achieve a bias less than 0.01 for a sample size greater than 200 and the estimates with small D do this for a sample size greater than 500. For estimates of variances with small, medium and large D the bias becomes less than 1 for a sample size greater than 100. Whereas the estimates in proportion with medium D of sample size greater than 200

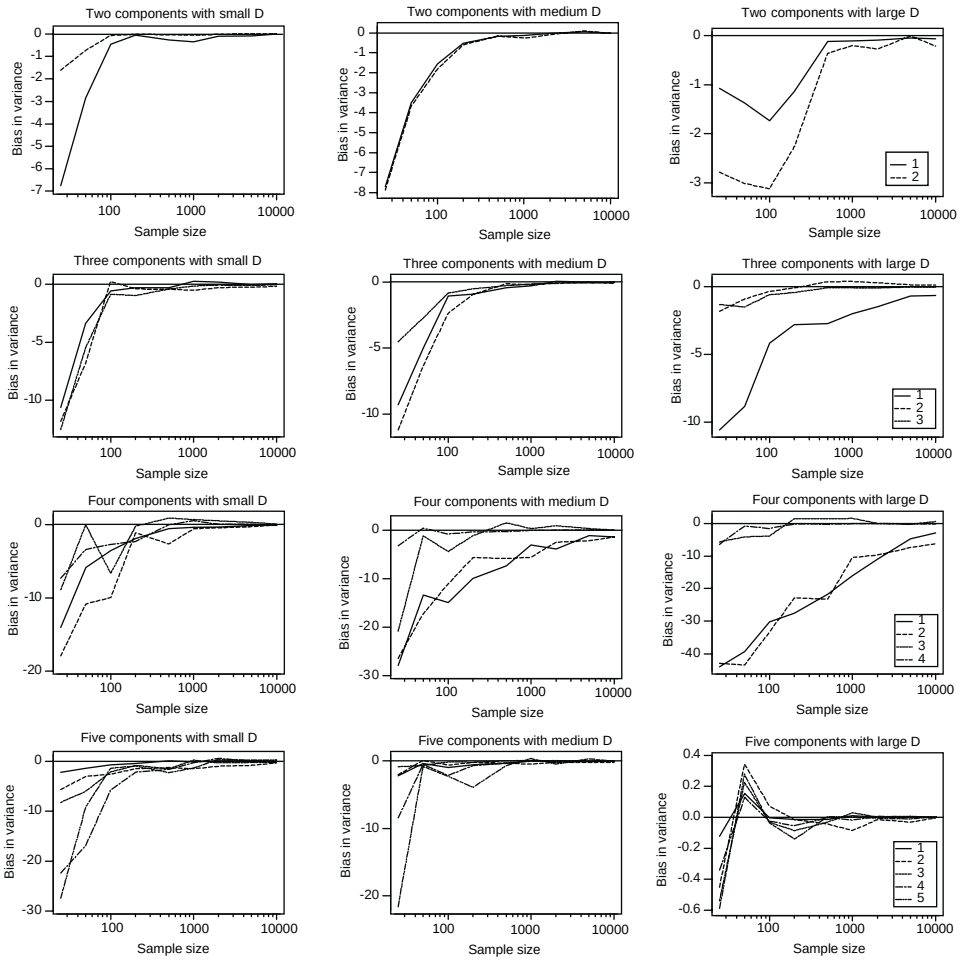

Fig. 3. Bias in variance for two to five components with the small, medium and large value of D.

and large D of sample size greater than 100 achieve a bias in proportion less than 0.01. In generality, the bias tends to approach zero as there is increase in the sample size. The pattern in bias is similar for two to five components.

For the MSE of mean, there is no difference in the pattern of convergence to zero for small, medium and large D. With a sample size greater than 100, the MSE of mean has value less than 0.1. For the MSE of variance, it tends to have a value less than 100 for small, medium and large D when the sample size greater than 500, 200, and 100, respectively. For the MSE of proportion, the components with small, medium and large value of D tends to have the MSE less than 0.001 for the sample size greater than 1000, 500 and 200, respectively.

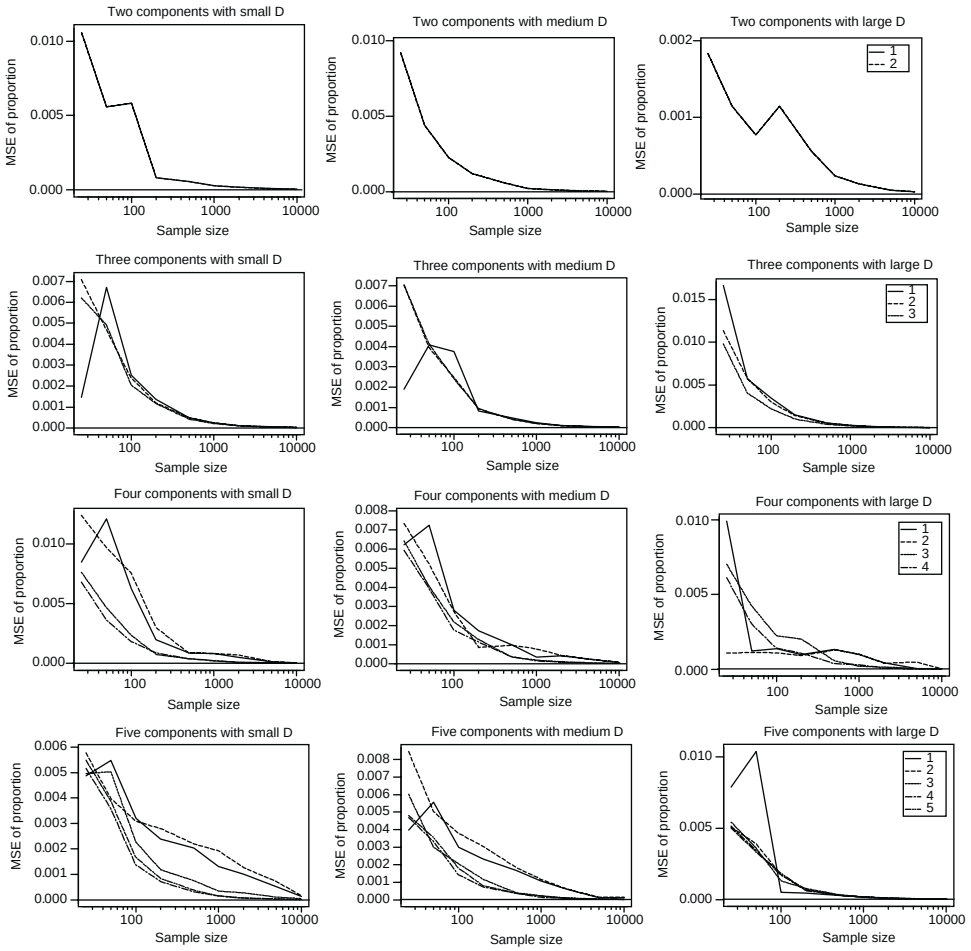

Fig. 4. Mean square error of proportion for two to five components with small, medium and large value of D.

5. Conclusions and discussion

For all situations investigated in this simulation study we have seen that the EM algorithm gives reasonable solutions of the score equations in an asymptotic unbiased sense. The value of the index D or the trace of covariance matrix of mixing the distribution of the normal distribution influences the pattern of convergence. For the data with medium valued D, reasonably small bias in mean and variance seem to occur for the sample size greater than 200.

As has been demonstrated the EM algorithm estimate seem to provide reasonable estimates of the parameter values. A special strategy has been used to find initial values. It remains up to future research how dependent these results are on the initial values chosen here, and if less favorable results are to be expected if other strategies are used.

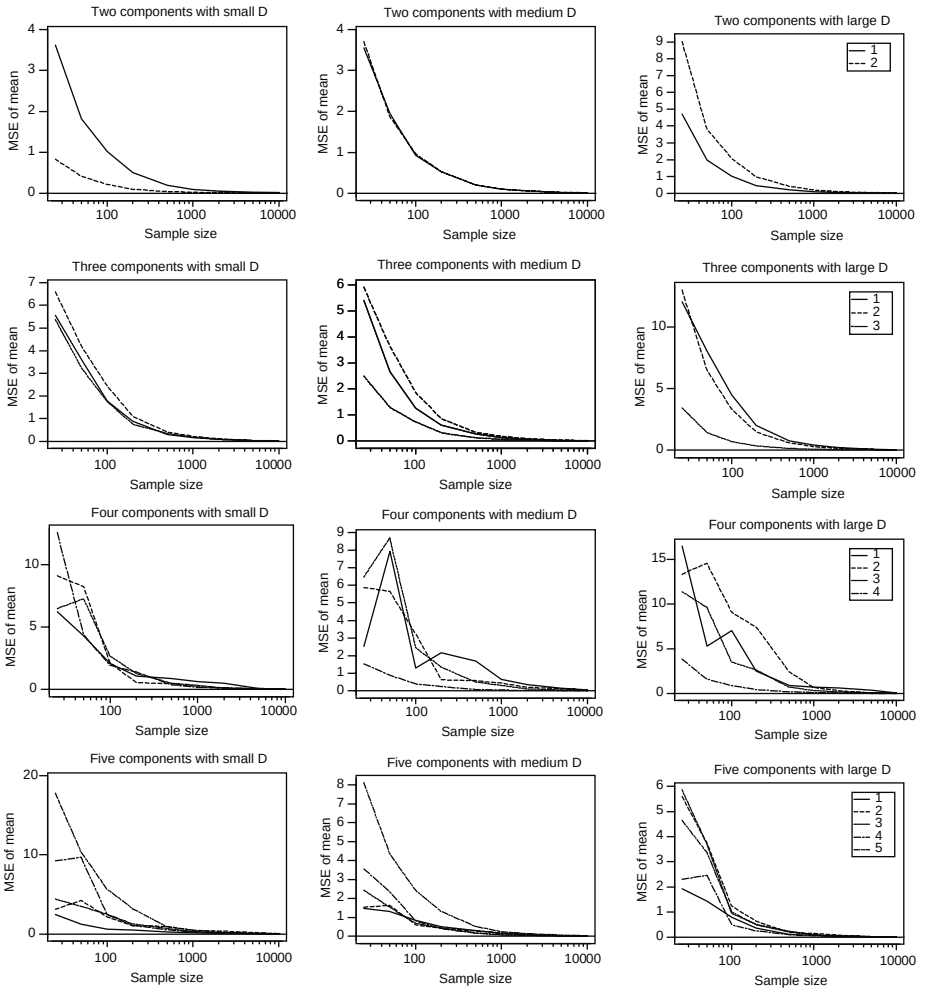

Fig. 5. Mean square error of mean for two to five components with small, medium and large value of D.

References

Behboodian, J., 1970. On a mixture of normal distributions. Biometrika 57, 215–217.

Böhning, D., 2000. Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others. CRC/Chapman & Hall, London.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion) J. Roy. Statist. Soc. B 39, 1–38.

McLachlan, G.J., Krishnan, T., 1997. The EM algorithm and Extensions. Wiley, New York.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

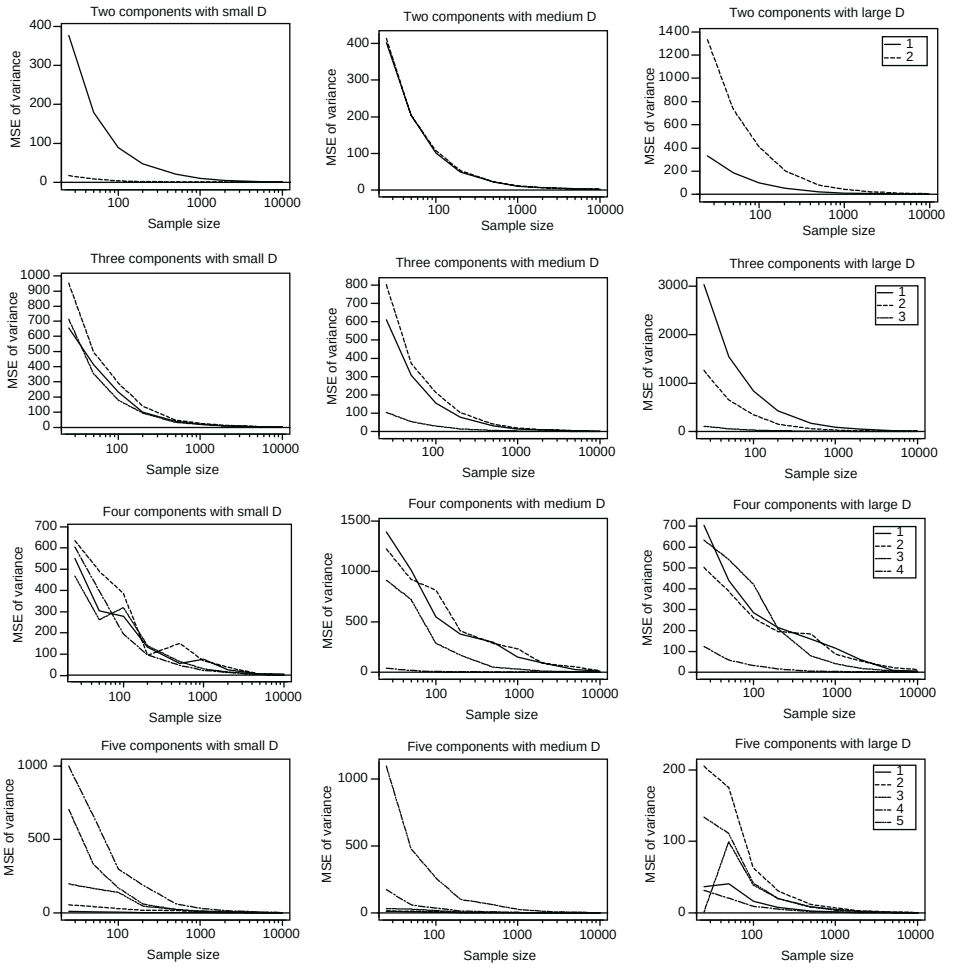

Fig. 6. Mean square error of variance for two to five components with small, medium and large value of D.

Meng, X.-L., 1997. The EM algorithm and medical studies: a historical link Statist. Methods Med. Res. 6, 3–23.

Redner, R.A., Walker, H.F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. 26, 195–239.

Wu, C.F., 1983. On the convergence properties of the EM algorithm of the EM algorithm. Ann. Statist. 11, 95–103.