International Journal of Computational and Theoretical Statistics
ISSN (2210-1519)
Int. J. Comp. Theo. Stat. 5, No. 2 (Nov-2018)

http://dx.doi.org/10.12785/ijcts/050204

Performance of the Stochastic EM Algorithm for Estimating Mixture Parameters

Athanase Polymenis[1]

[1] *Department of Economics, University of Patras, Patras, Greece*

Received April 18, 2018, Revised July 1, 2018, Accepted August 5, 2018, Published November 1, 2018

Abstract: The aim of the present study is to evaluate the performance of the Stochastic EM (SEM) algorithm for estimating parameters in finite mixture distributions. For that purpose some simulation exercises which are based on the implementation of this algorithm are presented and they provide encouraging results. As a consequence it seems that SEM can offer a useful alternative for overcoming some well-known difficulties arising when the number of mixture components is itself an unknown parameter.

Keywords: Mixture Distribution, Stochastic EM, Simulation

1. Introduction

Assume that X is a random vector which has a distribution with probability density function of the form $f(x) = \sum_{k=1}^{K} p_k f(x,\theta_k)$, where $f(x,\theta_k)$ is the density function of the k–th subpopulation/ component, p_k is the probability of belonging to the k–th subpopulation/component and is usually called mixing weight, with $0 < p_k < 1$ and $\sum_{k=1}^{K} p_k = 1$. We then say that the distribution of X is a finite mixture with K components. The aim of the present paper is to estimate the parameters of $f(x)$ using the Stochastic EM (SEM) algorithm. This algorithm has been well documented in earlier literature (see Celeux and Diebolt, [4,5,6], and Celeux, Chauveau and Diebolt, [3]). More recent research reveals the usefulness of the algorithm in diverse complex situations where mixture models are involved (see Bordes and Chauveau [2], and Teimouri, Rezakhah and Mohammadpour [10]). From a theoretical point of view this stochastic method can be regarded as a modified version of the well known EM algorithm (Redner and Walker, [9]), and was initially designed in order to circumvent some typical problems appearing when EM is used for estimating parameters in finite mixture distributions. Furthermore the method provides a useful tool for estimating the (unknown) number of components K. This is a nice feature of SEM keeping in mind that in a mixture context a major problem arises when testing for the number of components because of a breakdown in the regularity assumptions pertaining to the likelihood ratio test statistic λ , and as a result, $-2\log(\lambda)$ does not have any longer its usual asymptotical χ^2 distribution under the null hypothesis; a simple example about that is presented in Titterington, Smith, and Makov [11]. Thus estimating the number of components is a difficult problem, sometimes called “the mixture problem”, and there is a rich literature on different ways for tackling it. Some ways to overcome this problem are provided by making use of the so-called algorithmic methods, like for example the SEM algorithm which is the method of interest in the present paper. In [3,4,5] extensive simulations were used in order to evaluate the performance of SEM (see [4]) and compare it to that of some other algorithms like for example EM and some other EM driven methods (see [3,5]). In a slightly different direction the purpose of the present study is to also evaluate the SEM performance for estimating parameters in mixtures, but it is more focused on highlighting basic strengths of this algorithm on the basis of the investigation of some particular experimental cases.

E-mail address: athanase@econ.upatras.gr

2. THE SEM ALGORITHM

2.1. *Procedure*

Although as aforementioned this method has already been described in numerous papers we also present it in the sequel mainly for reasons of clarity. The idea underlying the methodology of SEM is to insert a stochastic step (S-step) between the expectation step (E-step) and the maximization step (M-step) of EM. More specifically, the procedure is as follows.

Considering the number of components to be unknown, define an upper bound K for this number and a threshold $c(N)$ lying between 0 and 1 (where by N we denote the sample size). The SEM iteration $q^r \rightarrow q^{r+1}$ is E-step: for $k = 1, \dots, K$ and $i = 1, \dots, N$ compute $t_k^r(x_i) = p_k^r\ f(x_i, \theta_k^r) / \sum_{j=1}^{K} p_j^r\ f(x_i, \theta_j^r)$.

S-step: for every observed x_i ($i = 1, \dots, N$) draw the pseudo-completed sample $y_i = \left(x_{i\,,}z(x_i)\right)$ by replacing each missing quantity $z(x_i)$ by a value drawn at random according to the probabilities $t_k^r(x_i)$, $k = 1, \dots, K$. This amounts to drawing a single multinomial observation $z^r(x_i)$=($z_k^r(x_i)$, $k = 1, \dots, K$) with probabilities $t_k^r(x_i)$, $k = 1, \dots, K$. The realizations $z^r(x_i)$ define a partition $P^r = \{P_1^r, \dots, P_k^r\}$ of the observed sample $x_1, \dots, x_n$, with P_k^r={$x_i | (z_k^r(x_i) = 1$}. If card(P_k^r)< $Nc(N)$ the algorithm is re-initialized.
M-step: compute the maximum likelihood estimates q^{r+1} using the pseudo-completed sample constructed at the S-step. This procedure amounts to computing $p_k^{r+1} = (1/N) \sum_{i=1}^{N} (z_k^r(x_i))$ and estimating the θ_k's.

Note that according to theory successive SEM iterates constitute an ergodic Markov chain (see Celeux and Diebolt [5,6]).

2.2. *Basic Properties*

We now summarize in brief some nice properties of SEM. We first remark that this algorithm was designed in order to overcome some drawbacks of EM for estimating parameters in mixtures. In relation to the problem of estimating numbers of components it is interesting to note that SEM allows for misspecifications of these numbers; indeed one need only know an upper bound of this number (see Celeux and Diebolt, [6]). Note that the S-step deletes a nice feature of EM, namely that the observed likelihood is increased at each iteration, but on the other hand it allows SEM to avoid saddle points and local maxima (in contrast to EM), and thus this algorithm converges in general faster than EM. Note that the main result for local convergence concerning SEM relies on Theorem 1, appearing in [6]. The theorem states in brief that under some regularity assumptions and provided that the sample size N is large enough the sequence of iterates generated by the SEM algorithm will converge in distribution to a stationary normal distribution approximately concentrated around the MLE. We finally mention that analytic properties of the SEM algorithm with practical applications in mixture models can be found in Celeux and Diebolt [4,5]. Note that in the latter articles simulation results that show nice features of SEM were provided.

3. SIMULATION STUDIES

3.1. *The General Models*

We now focus on some particular mixture models of interest and we evaluate the performance of the SEM algorithm on the basis of two experiments as we now present. Concerning the first experiment we consider the estimation of parameters in a two-normal component mixture situation with common variance. Data arising from such models were generated, and we evaluated the method's performance using diverse sample sizes; also different starting values for the parameters under estimation (especially the weights) were used in order to investigate whether these values could affect the performance of the algorithm. In all our models the normal distribution is chosen as the underlying distribution, and mixture components are considered to have a common variance, since as mentioned in McLachlan and Peel [7], p. 176, "any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities with common variance". Concerning the second experiment we generated data from a single normal component to which we fitted a mixture of two normals. This application refers to the very important matter of testing between one and two normal components. Note that for cases where the (one-dimensional) component distributions are Gaussian, $\theta_k$takes the special form $\theta_k = (\mu_k, \sigma_k)$ – where μ_k and σ_k refer respectively to the mean and the standard deviation of the k -th component. In a mixture context, these parameters are estimated via the SEM algorithm using the following algebraic expressions in the one-dimensional case (see Celeux and Diebolt [5] for the general form of these expressions): at iteration $r + 1$,

$$\mu_k^{r+1} = \sum_{i=1}^{N} z_k^r(x_i)x_i / \sum_{i=1}^{N} z_k^r(x_i) \ , \qquad (1)$$

and

$$\sigma_k^{r+1} = \sqrt{\sum_{i=1}^{N} z_k^r(x_i)(x_i - \mu_k^{r+1})^2 / \sum_{i=1}^{N} z_k^r(x_i)} \ . \quad (2)$$

(Note that in all our experiments we consider a known common variance for the components and thus equation (2) is irrelevant).

Finally, the mixing weight is estimated as reported in the previous section, i.e.

$$p_k^{r+1} = (1/N) \sum_{i=1}^{N} (z_k^r(x_i)) \ . \qquad (3)$$

3.2. Computing Issues

Concerning implementation issues for the aforementioned experiments, note that since as mentioned by Celeux, Chauveau, and Diebolt [3], p. 296, "SEM does not provide directly a pointwise estimate" , we used the simulation procedure proposed by these authors, pp. 296-297. That is, we choose to run a total number of 100 iterations for SEM, where a "warm-up" step of 75 iterations is first performed (in order to reach the stationary regime of SEM), and then we average the estimates over the remaining 25 iterations. The mean SEM estimates obtained from this procedure in the next sections are denoted by the symbol $\overline{(.)}$.

We also report that Fortran 77 was used for programming purposes. The Nag library was used for generating pseudorandom numbers.

3.3. First Experiment

The experiment concerns Model (1) which is a mixture of the form 1/2N(0,1)+1/2N(4,1), i.e. has two normal components, and a common variance (and thus corresponding data are generated from 1/2N(0,1)+1/2N(4,1)). The parameters to be estimated are the mixing weights and the means. Note that this model has also been used in an earlier paper concerning the stochastic EM algorithm, in a somehow different context (see Polymenis [8]). The model was chosen to have well-separated components in order to avoid problems associated with severely overlapping components like those reported in Section 6 of [3] for mixture *M1*. In the context of our experiment four sample sizes and different starting values of the parameters are used for the fitted models in five simulation exercises and are presented in Table 1. Furthermore consecutive SEM iterates are obtained using equations (1) and (3) and the parameter means, computed as reported in the previous section, are reported in Table 1. The method works well (as expected) for very large sample sizes (500 and 1000) and starting values for the parameters under estimation which are near the true parameter values. For a medium sample size of 200 and starting values not so near as for the two previous models, especially those concerning the mixing weights, the performance of SEM is quite good as well. For the rather small sample size (N = 100) and starting values $\mu_1 = 0.975$, $\mu_2 = 4.67$, $p_1 = 0.69$, almost equal to those used for N = 200, the performance of SEM is also fine, with estimated means near the true means and estimated weights only slightly further from the true weights than were corresponding estimated weights for N = 200. Thus decreasing the sample size from 200 to 100 does not seem to substantially affect the performance of the algorithm. Furthermore it is interesting to remark how well the method works for the case where the sample size is as before (N = 100) and the starting values are far from the true parameter values especially those concerning the mixing weights ($\mu_1 = 2.114$, $\mu_2 = 6.187$, $p_1 = 0.99$). Although the fitted model can be considered to be almost a single normal component (with one outlier), SEM manages to recover the true model. This result also shows the strength of the SEM algorithm since, as aforementioned, SEM detects the true model specification only when the fitted model includes a number of components which is at least equal to that of the true model; this practically means that in case an exact single normal component were fitted it would have been impossible for the algorithm to detect a two-component mixture.

Table 1. Initial and estimated parameters. The true model is 1/2N(0,1)+1/2N(4,1).

$N = 100$	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
$\mu_1 = 2.114$ $\overline{\mu_1} = 0.048$	$\mu_1 = 0.975$ $\overline{\mu_1} = 0.048$	$\mu_1 = 1.036$ $\overline{\mu_1} = 0.135$	$\mu_1 = 0.277$ $\overline{\mu_1} = 0.048$	$\mu_1 = 0.416$ $\overline{\mu_1} = 0.069$
$\mu_2 = 6.187$ $\overline{\mu_2} = 4.021$	$\mu_2 = 4.67$ $\overline{\mu_2} = 4.02$	$\mu_2 = 4.66$ $\overline{\mu_2} = 4.01$	$\mu_2 = 4.18$ $\overline{\mu_2} = 3.978$	$\mu_2 = 4.214$ $\overline{\mu_2} = 3.96$
$p_1 = 0.99$ $\overline{p_1} = 0.48$	$p_1 = 0.69$ $\overline{p_1} = 0.48$	$p_1 = 0.71$ $\overline{p_1} = 0.49$	$p_1 = 0.573$ $\overline{p_1} = 0.518$	$p_1 = 0.59$ $\overline{p_1} = 0.512$
$p_2 = 0.01$ $\overline{p_2} = 0.52$	$p_2 = 0.31$ $\overline{p_2} = 0.52$	$p_2 = 0.29$ $\overline{p_2} = 0.51$	$p_2 = 0.427$ $\overline{p_2} = 0.482$	$p_2 = 0.41$ $\overline{p_2} = 0.488$

Letters without the symbol $\overline{(.)}$ are initial parameters. Letters with the symbol $\overline{(.)}$ are estimated mean parameters using SEM.

3.4. *Second Experiment*

We now focus on our second experiment, that is, the case where data arise from a standardized normal model to which a mixture of two normal components is fitted.

Evidence concerning the performance of SEM is provided on the basis of two simulation exercises. More specifically let true Model (2) be a normal N(0,1) model (and thus data are generated from a standardized normal distribution) to which mixtures with two normal components of the form 0.606N(-0.365,1)+0.394N(0.491,1), which we call Model $(2)_1$, and 0.934N(-0.077,1)+0.066N(0.672,1), which we call Model $(2)_2$, are fitted. Like for Model (1) parameters under estimation are the mixing weights and the means. Note that Model $(2)_1$ has also been used in an earlier paper concerning the stochastic EM algorithm, in a somehow different context (see Polymenis [8]). These models were included into our analysis because although they are not standardized normals their shapes approximate that of N(0,1), as we now explain. Models $(2)_1$ and $(2)_2$ are both poorly separated (respective Mahalanobis distances are 0.856 and 0.749) and hence it is difficult to graphically distinguish their component densities (see Mc Lachlan and Peel, [7], p. 9). Also parameters involved in Model $(2)_2$ indicate that it has a shape close to that of a standardized normal since $(2)_2$ can be approximately considered as a standardized normal plus a few outliers whereas this is not so obvious for Model $(2)_1$. Concerning the latter model one can easily compute basic moments and compare them to corresponding ones from a standardized normal in order to check whether there is any similarity in their shapes. For this purpose we computed these moments for Model $(2)_1$ using results of proposition 2.2.1. of Wang [12] and found mean= -0.0277, standard deviation= 1.084, skewness= 0.025 and kurtosis=2.96, which are close to mean=0, standard deviation=1, skewness=0, and kurtosis=3 (the theoretical moments of N(0,1)) and thus the shape of the model distribution can be considered to be close to that of N(0,1). Our goal is then to investigate the performance of the SEM algorithm in these situations. Also note that in this experiment a large sample size of 500 is used in order to improve the performance of SEM. Like for the previous experiment equations (1) and (3) are used and corresponding parameter means are duly computed. Results are as follows. For the first simulation exercise, i.e., when Model $(2)_1$ is fitted, the estimated (mean) parameters of interest obtained from equations (1) and (3) are $\overline{p_1} = 0.922$, $\overline{\mu_1} = 0.00233$, $\overline{p_2} = 0.078$, $\overline{\mu_2} = -0.394$. These results point towards an approximate single component model, since $\overline{p_1}$ is much closer to 1 than were the starting value for p_1 (0.606), and $\overline{\mu_1}$ very close to the true mean 0. Furthermore, by results of proposition 2.2.1. of Wang [12], moments obtained for the estimated model are mean= -0.0286 , standard deviation= 1.0113, skewness=-0.003734, and kurtosis=3.00098, and thus are much closer to corresponding theoretical moments from N(0,1) than were initial moments of Model $(2)_1$. Concerning the second simulation exercise, i.e., when Model $(2)_2$ is fitted, corresponding (mean) estimates for the parameters of interest obtained from SEM are $\overline{p_1} = 0.998$, $\overline{\mu_1} = -0.0243$, $\overline{p_2} = 0.002$, $\overline{\mu_2} = -2.2$. These results show the good performance of the SEM algorithm since $\overline{p_1}$ is much closer to 1 than were its starting value (0.934), even though the latter value was also quite large. The second component has almost disappeared and thus a single component model is clearly indicated by SEM. Furthermore $\overline{\mu_1}$ is closer to the true mean value 0 than were the initial value (-0.077). Finally remark that SEM estimates obtained for Model $(2)_2$ are slightly more accurate than those for Model $(2)_1$.

4. CONCLUSION

The performance of the SEM algorithm was evaluated in the present paper. Our simulation results concerning the first experiment and presented in Table 1 indicate that when data are generated from a mixture with two normal components and a reasonable degree of components separation SEM shows in general a nice performance in detecting the true parameter values. The algorithm performs well even for an extreme case, presented in Table 1, where the mixing weight of the first fitted component is very close to 1, the fitted component means are away from the true means and the sample size is rather small. Concerning the second experiment, that is the case where data are generated from a standardized normal distribution, the fitted models are two-component mixtures which have quite different parameters but whose shapes resemble that of a single standardized normal component. This is a particular characteristic we included in order to assess its impact on the method since as far as we know it has not been investigated in previous literature concerning the SEM algorithm. Our simulation results show that after implementation of SEM the second component has almost disappeared and the estimated mean of the first component is close to 0 for both simulation exercises. This nice performance of SEM is also in accordance with results presented in [4,5] regarding the assessment of the algorithm in case the model is overfitted. We finally note that the method seems to work well for both experiments used in this paper independently of the starting values. On the other hand it is understood that this result is obtained within the limited framework of our study; more practical experiments would be needed in order to have a better insight about the role of starting values. However our results agree with the statements concerning starting values for SEM provided by Celeux and Diebolt [5, p. 10] and by Baudry and Celeux [1, p. 6], and with the results of [4,5].

ACKNOWLEDGMENT

The author is grateful to an anonymous referee for his useful suggestions that helped improve this text.

REFERENCES

[1] J. P. Baudry and G. Celeux, "EM for mixtures - Initialization requires special care," hal-01113242, 2015. (https://hal.inria.fr/hal-01113242)

[2] L. Bordes and D. Chauveau, "EM and Stochastic EM algorithms for reliability mixture models under random censoring," hal-00685823v2, 2012. (https://hal.archives-ouvertes.fr/hal-00685823v2)

[3] G. Celeux, D. Chauveau, and J. Diebolt, "Stochastic versions of the EM algorithm: an experimental study in the mixture case," J. Statist. Comput. Simul., vol. 55, pp. 287-314, 1996.

[4] G. Celeux and J. Diebolt, "Reconnaissance de mélange de densite et classification. Un algorithme d'apprentissage probabiliste: l'algorithme SEM," RR-0349, INRIA, (inria-00076208), 1984. (https://hal.inria.fr/inria-00076208)

[5] G. Celeux and J. Diebolt, "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem,"Computational Statistics Quarterly, vol. 2, pp. 73- 82, 1985.

[6] G. Celeux and J. Diebolt, "A random imputation principle: the stochastic EM algorithm," RR-0901, INRIA, (inria-00075655), 1988. https://hal.inria.fr/inria-00075655)

[7] G. McLachlan and D. Peel, Finite Mixture Models. Wiley: New York, 2000, pp. 9, 176.

[8] A. Polymenis, "A note on a validity test using the stochastic EM algorithm in order to assess the number of components in a finite mixture model," Statistics, vol. 42, pp. 261-274, 2008.

[9] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," SIAM Review, vol. 26 , pp. 195-239, 1984.

[10] M. Teimouri, S. Rezakhah, and A. Mohammadpour, "EM algorithm for symmetric stable mixture model," Communications in Statistics - Simulation and Computation, vol. 47(2), pp. 582-604, 2018.

[11] D. M. Titterington, A. F. M. Smith, and U. E. Makov, Statistical Analysis of Finite Mixture Distributions. Wiley: New York, 1985, pp. 152-153.

[12] J. Wang, “Generating daily changes in market variables using a multivariate mixture of normal distributions,” in Proceedings of the 2001 Winter Simulation Conference, B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, Eds., 2001, pp. 283-289.