Statistics and Computing (1999) **9**, 303–307

Improving the EM algorithm for mixtures

DIMITRIS KARLIS and EVDOKIA XEKALAKI

Department of Statistics, Athens University of Economics and Business, 76 Patision Str, 10434, Athens, Greece

Received April 1998 and accepted March 1999

One of the estimating equations of the Maximum Likelihood Estimation method, for finite mixtures of the one parameter exponential family, is the first moment equation. This can help considerably in reducing the labor and the cost of calculating the Maximum Likelihood estimates. In this paper it is shown that the EM algorithm can be substantially improved by using this result when applied for mixture models. A short discussion about other methods proposed for the calculation of the Maximum Likelihood estimates are also reported showing that the above findings can help in this direction too.

Keywords: EM algorithm, k-finite mixtures, normal mixtures, Poisson mixtures, semiparametric maximum likelihood

1. Introduction

Mixture models are widely used to describe inhomogeneous populations in a variety of fields of statistical applications including biological applications. Since inhomogeneity is a rather common fact in biological populations, mixture models are useful devices for its description. A well known example in fishery studies concerns the modeling of the weight of fish. Since the weight of a fish depends on its age, the presence of fish of different ages in the sample leads to the rejection of the simple homogeneous model. In this situation, a mixture model is more appropriate. Several other examples and applications of mixture models can be found. During the last few years, computers made possible the development of efficient algorithms that made the estimation of such models an easy task.

A k-finite mixture of a distribution defined by the probability density function $f(x \mid \theta)$ is defined by the density

$$g(x) = \sum_{j=1}^{k} p_j f(x \mid \theta_j) \tag{1}$$

where $p_j > 0$ for $j = 1, \ldots, k$; $\sum_{j=1}^{k} p_j = 1$ are the mixing proportions and θ_j are the parameters for each subpopulation. The mixing proportion p_j can be regarded as the probability that a randomly selected observation belongs to the j-th subpopulation. The parameters θ_j can be vector valued. This is a useful model to describe inhomogeneous populations that can be thought of consisting of k subpopulations. Each subpopulation has a distribution of the same parameteric form with varying parameter value. The distribution which assigns positive probability p_j at the point θ_j, $j = 1, \ldots, k$, is referred to as the mixing distribution.

Given a random sample $X_1, X_2, \ldots, X_n$, the likelihood function L is given by

$$L = \prod_{i=1}^{n} g(x_i) = \prod_{i=1}^{n} \left(\sum_{j=1}^{k} p_j f(x_i \mid \theta_j) \right).$$

The logarithm of the likelihood is therefore given by

$$\ell = \log L = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} p_j f(x_i \mid \theta_j) \right). \tag{2}$$

In order to obtain the Maximum Likelihood Estimates (MLE) for the k-finite mixture model, we have to differentiate (2) with respect to the parameters and to equate the result to zero. Thus, the estimating equations are:

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{n} \frac{p_j}{g(x)} \frac{\partial f(x_i \mid \theta_j)}{\partial \theta_j} = 0, \quad j = 1, \ldots, k, \tag{3}$$

$$\frac{\partial l}{\partial p_j} = \sum_{i=1}^{n} \frac{f(x_i \mid \theta_j) - f(x_i \mid \theta_k)}{g(x_i)} = 0, \quad j = 1, \ldots, k-1. \tag{4}$$

The system of equations (3) and (4) must be solved to obtain the MLE. In the next section, it is shown that if $f(x \mid \theta)$ belongs to the one-parameter exponential family,

the ML estimators satisfy the first moment equation. Some of the best known distributions belong to this family and hence the results can be applied to many cases where finite normal mixtures, finite exponential mixtures, finite Poisson mixtures, among other models are appropriate. In Section 3 we describe how this result can be used to substantially improve the speed of the well known EM algorithm for mixtures, providing some simulation results for the cases of finite normal and finite Poisson mixtures. We also provide a discussion on how some other algorithms designed to find the MLE can be improved using the results of Section 2.

2. The main result

Let us consider that the density $f(x \mid \theta)$ comes from the one-parameter exponential family, i.e. that $f(x \mid \theta)$ can be written in the form

$$f(x \mid \theta) = \exp[\theta x c + h(x) - k(\theta)], \tag{5}$$

where c is some constant and the functions $h(x)$ and $k(\theta)$ depend only on x and θ respectively.

From the estimating equations for the general finite mixture model given in (3) and (4) we obtain, by multiplying the ith equation in (4) by p_j, $j = 1, 2, \ldots, k$, and adding the resulting equations

$$\sum_{i=1}^{n} \frac{f(x_i \mid \theta_k)}{g(x_i)} = n. \tag{6}$$

On the other hand, since from (4)

$$\sum_{i=1}^{n} \frac{f(x_i \mid \theta_j)}{g(x_i)} = \sum_{i=1}^{n} \frac{f(x_i \mid \theta_k)}{g(x_i)}, \quad j = 1, \ldots, k,$$

it may be concluded that the maximum likelihood estimates satisfy

$$\sum_{i=1}^{n} \frac{f(x_i \mid \theta_j)}{g(x_i)} = n, \quad j = 1, \ldots, k. \tag{7}$$

Also, from (3) and (5) it follows that

$$\sum_{i=1}^{n} \frac{f(x_i \mid \theta_j)}{g(x_i)} \big(x_i - \mu(\theta_j)\big) = 0, \quad j = 1, \ldots, k. \tag{8}$$

Then, setting $w_{ij} = f(x_i \mid \theta_j)/g(x_i)$, $i = 1, \ldots, n$, $j = 1, \ldots, k$ equation (8) can be written as

$$\sum_{i=1}^{n} w_{ij}\big(x_i - \mu(\theta_j)\big) = 0, \quad j = 1, \ldots, k. \tag{9}$$

If we consider the mean value reparameterization for the density $f(x \mid \theta)$ and solve for the mean value parameters we obtain by combining (7) and (9)

$$\mu(\theta_j) = \frac{\sum_{i=1}^{n} w_{ij} x_i}{n}, \quad j = 1, \ldots, k. \tag{10}$$

As is well known, the mean of a mixture is the weighted mean of the means of all components weighted by the mixing proportions. Then, from (1) and (10) the estimator of the mean of the finite mixture is

$$\sum_{j=1}^{k} p_j \mu(\theta_j) = \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} w_{ij} x_i}{n} p_j = \frac{\sum_{i=1}^{n} \frac{\sum_{j=1}^{k} p_j f(x_i \mid \theta_j)}{g(x_i)} x_i}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x},$$

namely the sample mean.

Hence, the MLE of the mean value parameter of a k-finite mixture from the one-parameter exponential family coincides with the first moment. An alternative proof of this result which seems to have passed unnoticed was given earlier by Lindsay (1981). The above result is true also for members of the power series family of distributions (see for example Johnson *et al.*, 1992). Sprott (1983) showed that this result holds for the convolution of two power series distributions as well as for compound (or generalized) distributions of members of the power series family. A generalization of the power series family shares the same property as Kemp (1986) showed. It is interesting that some of the most well known distributions belong to the one-parameter exponential family, like the Poisson, the normal, the exponential, the gamma and other distributions. For many of them the parameter θ represents the mean of the distribution. Behboodian (1970) has shown a similar result for finite normal mixtures. The above results hold also in the semi-parametric case where the number of support points is not known a priori (see Lindsay, 1995).

In the next section, the above result is used for improving the EM algorithm.

3. Improving the EM algorithm for finite mixtures and other applications

From the previous section, it becomes obvious that the estimating procedure can be simplified if one of the equations in (3) or (4) is replaced by the first moment equation. For example, the EM algorithm proposed by Hasselblad (1969) to deal with the ML estimation in mixture models is an iterative algorithm using the above equations. The EM algorithm can be described as follows:

E-step: With the current estimates p_j^{old} and $\mu(\theta_j^{\text{old}})$ calculate

$$w_{ij} = p_j^{\text{old}} f(x_i \mid \theta_j^{\text{old}})/g(x_i), \quad i = 1, \ldots, n, \quad j = 1, \ldots, k.$$

M-step: Obtain the new estimates of the parameters $\mu(\theta_j)$ and p_j from

$$\mu(\theta_j^{\text{new}}) = \frac{\sum_{i=1}^{n} w_{ij} x_i}{\sum_{i=1}^{n} w_{ij}} \quad \text{and} \quad p_j^{\text{new}} = \frac{\sum_{i=1}^{n} w_{ij}}{n}, \quad j = 1, \ldots, k.$$

Then, go back to the E-step replacing the old values with the new ones from the M-step.

The iterative scheme terminates when some condition (indicating convergence) is satisfied. We can verify easily that the above scheme always satisfies the requirement for the first moment.

The weights w_{ij} are the posterior probabilities of the observation X_i to belong to the j-th subpopulation. This representation is more concrete with respect to the general form of the EM algorithm for missing value problems or problems which can be considered as 'missing value problems'. The latter include the case of finite mixtures where complete knowledge of the subpopulations to which the observations belong would make the estimation straightforward. Since we do not know the subpopulation to which a particular observation belongs, we estimate it via the posterior probabilities. Hasselblad (1969) and Behboodian (1970) independently introduced this iterative algorithm, prior to the general derivation of the EM algorithm by Dempster *et al.* (1977).

The EM algorithm for finite mixtures is widely applicable because of its simple and easily programmable form. However, it has the disadvantage of slow convergence and high dependence on the initial values, but this seems to be a common problem with iterative algorithms if the likelihood equation has multiple roots (see, for example McLachlan and Krishnan, 1997). Thus, since the EM algorithm may stop with a local maximum which is not global, several initial values must be used. This makes the algorithm very time demanding. Improvements have been proposed in three different directions. Bohning *et al.* (1994) propose an easier method for detecting the convergence of the algorithm saving thus iterations. Fruman and Lindsay (1994) recommended the use of efficient initial values, namely the use of the moment estimates as initial values for the EM algorithm. Lange (1995) proposed quasi-Newton acceleration, while Aitkin and Aitkin (1996) proposed that we can speed up the convergence alternating EM iteration with Gauss-Newton iterations.

Our results of the previous section can also serve as a basis for improving the EM algorithm for finite mixtures. Our approach can be combined with the above mentioned methods. In this case, the gain in computational time will be maximized.

So, using the results of Section 2, at each iteration the number of estimated parameters is reduced by one as one parameter can be estimated by the first moment equation. The gain in computing time is large as shown in Tables 1–3 for small values of k. If we look at the iterative scheme described above, we can see that we can avoid calculating $\mu(\theta_k)$ and this is equivalent to reducing the calculations involved for obtaining the new parameters by almost $100/(2k-1)\%$. In fact, the gain is less because of the cost for some additional calculations in each iteration. It is also interesting that the gain is expected to be larger in the case of discrete distributions, like the Poisson or the binomial distributions. This is so, because in the case of discrete distributions we can avoid exhausting summations by multiplying with the observed frequencies.

The examples that follow illustrate the gain in time using the method suggested.

Example 1: (Finite Poisson mixtures.) Consider the case of finite Poisson mixtures. In this case, we assume that $f(x \mid \theta) = \exp(-\theta)\theta^x/x!$. In order to examine the gain, a small simulation comparison was carried out. For $k = 2$, 100 samples of given sample size $n(n = 50, 100, 250, 500)$ were simulated for each distribution with parameter vectors $(p_1 = p,\ \theta_1 = 1,\ \theta_2)$. The times required for the ML estimation via the EM algorithm using both the general EM algorithm and the improved EM algorithm discussed above were calculated. The entries in Table 1 represent relative times, i.e. ratios of the times under the improved EM algorithm divided by the corresponding times under the standard EM algorithm. The time spent for simulating the samples was subtracted from both the numerator and the denominator. We tried to minimize the computing time for some auxiliary procedures like the terminating conditions. For each sample we stopped running the algorithm after 50 iterations. All the calculations were carried with a PC with a Pentium microprocessor. The results of Table 1 clearly show that we can save almost 20% of the computing time for $k = 2$.

Table 2 contains the results for $k = 3$. The vectors of parameters were $(p_1, p_2 = 0.3,\ \theta_1 = 1,\ \theta_2 = 2,\ \theta_3)$. For each distribution, 100 samples of given sample size $n(n = 50, 100, 250, 500)$ were simulated and the times required for both methods were recorded. The entries are again ratios of the time under the improved EM algorithm divided by the time under the standard EM algorithm. We can also see an improvement on the required computational time near 15%.

Example 2: (Finite Normal mixtures.) Behboodian (1970) showed that for the case of normal mixtures with different variances, the second moment equation is also satisfied, i.e. the ML estimator of the variance coincides with the sample variance. So, in the case of normal mixtures, at each EM

Table 1. *Times for the improved EM relative to the standard EM for 2-finite Poisson mixtures ($k = 2$)*

	$p_1 = 0.25$			$p_1 = 0.50$			$p_1 = 0.75$		
θ_2	2	5	10	2	5	10	2	5	10
n									
50	0.822	0.803	0.797	0.821	0.800	0.799	0.831	0.815	0.806
100	0.823	0.800	0.792	0.812	0.799	0.797	0.789	0.808	0.800
250	0.809	0.795	0.792	0.813	0.796	0.795	0.819	0.803	0.794
500	0.813	0.793	0.791	0.812	0.752	0.794	0.809	0.820	0.793

Table 2. *Times for the improved EM relative to the standard EM for 3-finite Poisson mixtures ($k = 3$)*

	$p_1 = 0.25$			$p_1 = 0.50$			$p_1 = 0.75$		
θ_3	3	5	10	3	5	10	3	5	10
n									
50	0.869	0.859	0.853	0.868	0.863	0.857	0.875	0.870	0.866
100	0.866	0.860	0.853	0.866	0.861	0.854	0.871	0.862	0.858
250	0.863	0.857	0.851	0.862	0.857	0.851	0.863	0.859	0.854
500	0.862	0.856	0.850	0.859	0.855	0.850	0.863	0.858	0.851

iteration we can simplify the estimation of two parameters. Then, we only have to calculate the $3k - 3$ parameters, while the remaining 2 parameters can be easily obtained by equating the first two moments to their sample counterparts. Thus, we reduce the required computational effort almost by a factor of $2/(3k - 1)$. In practice, the gain is less than that because of the cost of some additional calculations at each iteration. For $k = 2$, we simulated 100 samples from several 2-finite normal mixtures. The gain is near 30 (we estimate 3 parameters instead of 5) as can be seen in Table 3, for selected parameter vectors $\boldsymbol{\theta} = (p_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ and varying sample sizes. The entries of Table 3 are again ratios of the computing times required under the improved EM algorithm divided by the corresponding computing times under the standard EM algorithm for the same samples. Again, we tried to minimize any auxiliary calculations, and comments similar to those from the Poisson case apply here too.

The above findings provide a useful insight into the ML estimation for finite mixtures when the number of support points is not known a priori (see Bohning, 1995). Some authors refer to such models as semiparametric models (see Lindsay and Roeder, 1995). In the semiparametric case, one tries to estimate the nonparametric ML estimate of the mixing distribution without knowledge of the number of support points. Such algorithms usually add one new support point at each step and try to determine the probability to be assigned to this point. Consider the Vertex Direction Method. This method adds a new support point at the point where the gradient function is maximized. Numerical techniques are required in order to find the probability assigned to this point. Since for every value of k, the first moment equation must be satisfied, the solution with k points should satisfy this condition too. Suppose that at this moment we have k points, say $\mu(\theta_j)$ with associated probabilities p_j, for $j = 1, \ldots, k$. Then, if $\bar{x}$ is the sample mean it holds that $\sum_{j=1}^{k} p_j \mu(\theta_j) = \bar{x}$. Thus, the new support point, say $\mu(\theta_{k+1})$, will be assigned a probability α whose value is such that the increase in the loglikelihood under the transition from the model with k points to the model with $k + 1$ points is maximized. (See, for example an interesting review for determining the value of α in Bohning, 1995). This procedure usually requires specific numerical methods. If the new point is $\mu(\theta_{k+1})$, condition $(1 - a) \sum_{j=1}^{k} p_j \mu(\theta_j) + a\mu(\theta_{k+1}) = \bar{x}$ ought to be satisfied as this equation is one of the estimating equations in ML estimation with $k + 1$ support points. In any other case, the increase in the loglikelihood will not have been maximized. Solving with respect to α we obtain $\alpha = 0$, which implies that the new support point is rejected. Any other choice of α would lead to a solution which is not an ML solution with $k + 1$ support points.

Table 3. *Times for the improved EM relative to the standard EM for 2-finite Normal mixtures ($k = 2$)*

Vector of parameters $(p_1, \mu_1 z, \mu_2, \sigma_1^2, \sigma_2^2)$	Sample size			
	$n = 50$	$n = 100$	$n = 250$	$n = 500$
(0.25, 0, −1, 1, 2)	0.707	0.731	0.720	0.724
(0.25, 0, 1, 1, 2)	0.731	0.706	0.747	0.693
(0.5, 0, −1, 1, 2)	0.722	0.725	0.729	0.730
(0.5, 0, 1, 1, 2)	0.719	0.728	0.725	0.722
(0.75, 0, −1, 1, 2)	0.711	0.717	0.723	0.708
(0.75, 0, 1, 1, 2)	0.729	0.732	0.735	0.727
(0.25, 0, −1, 1, 5)	0.728	0.730	0.718	0.726
(0.25, 0, 1, 1, 5)	0.726	0.731	0.736	0.731
(0.5, 0, −1, 1, 5)	0.729	0.733	0.729	0.734
(0.5, 0, 1, 1, 5)	0.731	0.728	0.727	0.730
(0.75, 0, −1, 1, 5)	0.724	0.729	0.731	0.710
(0.75, 0, 1, 1, 5)	0.720	0.724	0.717	0.721

The above discussion reveals that when using the simple Vertex Direction Method, if the solution with k support points maximizes the likelihood for k-support points, no other point will be added and the method will fail. A similar behavior is expected to be shown by other related methods.

Therefore, some EM iterations are needed in order to improve the likelihood for each k. We can also see that the support points should change between two successive steps (not necessarily all of them) as otherwise, the new support point will be dropped. This makes the use of some EM iterations very useful, but in fact it cancels the applicability of all such methods. The reason is that since we have to apply EM iterations for each value k, the algorithm reduces to one deriving the MLE of each value of k via the EM and simply checking if this is the global maximum using the results of Lindsay (1983). Other algorithms like, the Vertex Exchange method or the Intra-Simplex Direction method (see Bohning, 1995) show the same behavior.

4. Conclusions

It has been shown that one of the ML equations in finite mixtures of members of the one-parameter exponential family can be replaced by the first moment equation. This simplifies the procedure for the ML estimation and can save a lot of computational time by speeding up the EM

algorithm. It has also been revealed that other methods for MLE in the case of unknown k are not efficient and they have to be used with caution. For distributions with discrete density $f(x \mid \theta)$, the above scheme can be modified so as to be applied to minimum distance estimation for finite mixtures. The key tool is the first derivative of $f(x \mid \theta)$ which allows the estimates to be written in the form of a weighted mean. The above iterative scheme appropriately modified has already been applied to Minimum Hellinger Distance Estimation for finite Poisson mixtures (Karlis and Xekalaki, 1998). Extensions to multivariate finite mixtures may also be derived. For example, in the case of finite multivariate normal mixtures (see for example Day, 1969) the vector of means must coincide with the vector of sample means and therefore the calculations can be reduced considerably.

Acknowledgment

Dimitris Karlis is being supported by a scholarship from the State Scholarships Foundation of Greece.

References

Aitkin, M. and Aitkin, I. (1996) An hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing*, **6**, 127–130.

Behboodian, J. (1970) On a mixture of normal distributions. *Biometrika*, **56**, 215–217.

Bohning, D. (1995) A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference*, **47**, 5–28.

Bohning, D., Dietz, Ek., Schaub, R., Schlattman, P. and Lindsay, B. (1994) The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**, 373–388.

Day, N. E. (1969) Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.

Dempster, A. P., Laird, N. M. and Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **B39**, 1–38.

Fruman, W. D. and Lindsay, B. (1994) Measuring the relative effectiveness of moment estimators as starting values in maximizing mixture likelihoods. *Computational Statistics and Data Analysis*, **17**, 493–507.

Hasselblad, V. (1969) Estimation of finite mixtures from the exponential family. *Journal of the American Statistical Association*, **64**, 1459–1471.

Johnson, N., Kotz, S. and Kemp, A. (1992). *Univariate discrete distributions (2nd edition)*. Willey, New York.

Karlis, D. and Xekalaki, E. (1998) Minimum Hellinger distance Estimation for finite Poisson mixtures. *Computational Statistics and Data Analysis*, **29**, 81–103.

Kemp, A. W. (1986) Weighted discrepancies and Maximum Likelihood for discrete distributions. *Communications in Statistics*, **15**, 783–801.

Lange, K. (1995) A Quasi-Newton Acceleration of the EM algorithm. *Statistica Sinica*, **5**, 1–8.

Lindsay, B. (1981) Properties of the Maximum Likelihood Estimator of a Mixing Distribution. In G. P. Patil (ed) *Statistical Distributions in Scientific Work*, **5**, pp 95–109, Reidel, Boston.

Lindsay, B. (1983) The geometry of mixture likelihood. A general theory. *Annals of Statistics*, **11**, 86–94.

Lindsay, B. (1995) *Mixture Models: Theory, Geometry and Applications*. Regional Conference Series in Probability and Statistics, Vol 5, Institute of Mathematical Statistics and American Statistical Association.

Lindsay, B. and Roeder, K. (1995) A review of semiparametric mixture models. *Journal of Statistical Planning and Inference*, **47**, 29–39.

McLachlan, G. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. Willey Series.

Sprott, D. (1983) Estimating the parameters of a convolution by Maximum Likelihood. *Journal of the American Statistical Association*, **78**, 457–460.