HERCMA '98
E. A. Lipitakis (Editor)
1998, LEA
Printed in Hellas

AN IMPROVEMENT OF THE EM ALGORITHM FOR FINITE POISSON MIXTURES

Dimitris Karlis and Evdokia Xekalaki
Department of Statistics
Athens University of Economics and Business
76 Patision st, 10434, Athens, GREECE
email: karlis@stat-athens.aueb.gr, exek@aueb.gr

Abstract

Finite Poisson mixtures can be used in a variety of real applications to describe count data as they can describe situations where overdispersion relative to the simple Poisson model is present. They also admit a natural interpretation: the entire population is a mixture of k subpopulations each having a Poisson distribution giving rise to the k-finite Poisson distribution. Estimating the parameters of a k-finite Poisson mixture is not easy. However, the development of the EM algorithm for finite mixtures simplified the derivation of the maximum likelihood estimates. In this paper an improvement of the standard EM algorithm for finite Poisson mixtures is introduced. It is based on the result that one from the estimating equations for the Maximum Likelihood Estimates in the case of finite Poisson mixtures is the first moment equation. Hence, replacing one of the estimating equations by this simpler form can help us considerably in reducing the labour and the cost of calculating the MLE. Tables verifying the results are also given.

Keywords and phrases: Poisson mixtures; k-finite mixtures; EM algorithm; maximum likelihood estimation.

1. Introduction

Mixture models, finite or not, are widely used to describe inhomogeneous populations in a variety of fields of statistical applications as diverse as biological and actuarial applications. Since inhomogeneity is a rather common situation in biological populations, mixture models are useful devices for its description. A well known example in actuarial research concerns the modeling of the number of accidents of an insured driver. Since the driving ability and the exposure of a driver to external risk differs among the drivers, it seems reasonable to assume that the population is not homogeneous and thus the inhomogeneity can be represented by a mixture. Several other practical situations involving non-homogeneous populations can be modeled by mixtures upon similar assumptions that admit a physical interpretation in the context of the particular case. During the last few years computers made possible the development of efficient algorithms that made the estimation of such models a simple task.

The Poisson distribution plays a prominent role in describing count data that occur randomly. A basic assumption for the Poisson distribution is that the population is homogeneous. As already mentioned, inhomogeneity can be represented via mixtures of the Poisson distribution. These can be finite or not.

Finite Poisson mixtures are widely used in practice as, often, estimating the mixing distribution numerically one is restricted to estimating a finite number of mixing proportions. A k-finite mixture of Poisson distributions is defined as the distribution having probability function of the form

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

where $f(x|\theta) = exp(-\theta)\ \theta^x /x!$, *for* $x=0,1,...,$ $\theta>0$, i.e. the probability function of a Poisson distribution with parameter θ, and $p_i > 0$ for $i=1,..,k$ with $\sum_{j=1}^{k} p_j = 1$, are the mixing proportions. The latter can also be considered as the probabilities that an observation belongs to subpopulations 1,2,...,k respectively.

Given a random sample $X_1, X_2, ..., X_n$, the loglikelihood is given by

$$l = \sum_{i=1}^{n} \log(\sum_{j=1}^{k} p_j f(x_i|\theta_j)) \tag{2}$$

Denoting by $d(x)$ the frequency of the value x the loglikelihood in (2) can be rewritten as

$$l = \log L = \sum_{x=0}^{m} d(x) \log(\sum_{j=1}^{k} p_j f(x|\theta_j)) \quad , \tag{3}$$

where m is the largest value of the sample.

In the next section it is shown that the ML estimators satisfy the first moment equation.

2. The main result.

As usual, in order to find the MLE we need to equate all the partial derivatives of the loglikelihood with 0. Observing that, for the Poisson distribution, it holds that

$$\frac{\partial f(x|\theta)}{\partial \theta} = f(x-1|\theta) - f(x|\theta) \ ,$$

the estimating equations are

$$\frac{\partial l}{\partial \theta_j} = \sum_{x=0}^{m} d(x) \frac{p_j (f(x-1|\theta_j) - f(x|\theta_j))}{g(x)} = 0, \quad j=1,..,k \quad , \qquad (4)$$

$$\frac{\partial l}{\partial p_j} = \sum_{x=0}^{m} d(x) \frac{f(x|\theta_j) - f(x|\theta_k)}{g(x)} = 0, \qquad j=1,..,k-1 \qquad (5)$$

The system of equations (4) and (5) must be solved to obtain the MLEs.

From equations (4) we obtain

$$\sum_{x=0}^{m} \frac{d(x)}{g(x)} f(x|\theta_j) = \sum_{x=0}^{m} \frac{d(x)}{g(x)} f(x-1|\theta_j), \qquad j=1,..,k \qquad (6)$$

Also, multiplying the i-th equation in (5) by p_j, $j=1,..,k$ and adding the resulting equations we obtain

$$\sum_{x=0}^{m} \frac{d(x)}{g(x)} [g(x) - f(x|\theta_k)] = 0 \ , \qquad \text{or, equivalently,} \qquad \sum_{x=0}^{m} \frac{d(x)}{g(x)} f(x|\theta_k) = n \ . \qquad (7)$$

On the other hand, it follows from (4) that

$$\sum_{x=0}^{m} \frac{d(x)}{g(x)} f(x|\theta_j) = \sum_{x=0}^{m} \frac{d(x)}{g(x)} f(x|\theta_{kj}), \quad \text{j=1,...k} \qquad (8)$$

From (7) and (8) it may be concluded that the maximum likelihood estimates satisfy the equation

$$\sum_{x=0}^{m} \frac{d(x)}{g(x)} f(x|\theta_j) = n \ . \qquad (9)$$

Many authors (see, for example, Bohning, 1995) refer to the function in the right hand side of (9) as the gradient function, and they use it to check if the maximum is obtained. Bohning(1995) shows that the above conditions are necessary and sufficient for the θ_j 's to be MLE. Combining (6) and (9) it can be easily verified that

$$\sum_{x=0}^{m} \frac{d(x)}{g(x)} f(x-1|\theta_j) = n \ , \qquad j=1,2,...,k \ .. \qquad (10)$$

Adding equations (4) over j we obtain that

$$\sum_{x=0}^{m}\frac{d(x)}{g(x)}[g(x-1)-g(x)]=0\,, \qquad \text{or, equivalently}\,, \qquad \sum_{x=0}^{m}\frac{d(x)}{g(x)}g(x-1)=n\ . \tag{11}$$

i.e., the probability function of a k-finite mixture of Poisson distributions satisfies the same recurrence relationship as the probability functions of the Poisson components.

As is well known, $f(x-1|\theta)=f(x|\theta)\, x/\theta$. Then, (4) can be written as

$$\sum_{x=0}^{m}\frac{d(x)}{g(x)}f(x|\theta_j)(x-\theta_j)=0\,,$$

which setting $w_{xj}=f(x|\theta_j)/g(x)$ reduces to

$$\sum_{x=0}^{m}d(x)w_{xj}(x-\theta_j)=0\ .$$

Solving for the parameters, we obtain that

$$\theta_j=\frac{\sum_{x=0}^{m}d(x)w_{xj}x}{\sum_{x=0}^{m}d(x)w_{xj}}\ , \quad j=1,\dots,k \tag{12a}$$

Since from (9) the denominator is equal to n, (12a) becomes

$$\theta_j=\frac{\sum_{x=0}^{m}d(x)w_{xj}x}{n}\ , \quad \text{j=1,...,k} \tag{12b}$$

The above relationship implies that the ML estimators of the mean value parameters can be written as weighted sample means.

Suppose now that we have the MLE's for the parameters θ_j , $j=1,2,\dots k$. Then, from (1), it follows that the ML estimate of the mean of the finite Poisson mixture is

$$\sum_{j=1}^{k}\frac{\sum_{x=0}^{m}d(x)w_{xj}x}{n}p_j=\frac{\sum_{j=1}^{k}\sum_{x=0}^{m}d(x)w_{xj}xp_j}{n}=\frac{\sum_{x=0}^{m}\frac{d(x)}{g(x)}x\sum_{j=1}^{k}p_jf(x|\theta_j)}{n}=\frac{\sum_{x=0}^{m}d(x)x}{n}=\bar{x}\,,$$

i.e.. the sample mean.

Hence, the estimating equation leading to the ML estimate of the mean of a k-finite mixture of Poisson distributions coincides with the first moment equation. This is true also for members of the power series distribution (see, for example, Johnson *et al.*, 1992). Sprott (1983) showed that this is true 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. Finite Poisson mixtures belong to this family of distributions.

It becomes obvious, therefore, that the estimating procedure can be simplified if one of the equations in (4) and (5) is replaced by the first moment equation. For example, the EM algorithm proposed by Hasselblad (1969) to deal with ML estimation in mixture models, is an iterative algorithm using the above equations. The EM can be described as follows:

Starting with the current estimates $p_j^{old} and \theta_j^{old}$ calculate $w_{xj} = f(x|\theta_j)/g(x)$ and obtain the new estimates of the parameters using

$$\theta_j^{new} = \frac{\sum_{x=0}^{m} d(x) w_{xj} x}{n},$$

$$p_j^{new} = \frac{\sum_{x=0}^{m} d(x) p_j^{old} w_{xj}}{n}.$$

Then, go back and obtain the new values for the w_{xj} 's. The iterative scheme terminates when some condition is satisfied. We can verify easily that the above scheme always satisfies the requirement for the first moment.

Note that the representation of the weights w_{xj} is slightly different from that in the formal description of the EM algorithm where w_{xj} is defined as the posterior probability that, given the parameters, the observation with value x belongs to the j-th subpopulation *(j=1,. . . ,k)* . We adapt this representation since it enables the derivation of similar algorithms for other methods of estimation, as it has already been done by Karlis and Xekalaki (1998).

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. Thus, since the EM algorithm may stop at 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 a method for easier 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. Aitkin and Aitkin (1996) and Lange (1995) proposed alternating EM iterations and Gauss-Newton iterations.

So, 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 high for small values of k. Looking at the iterative scheme described above, it can be seen that calculating θ_k can be avoided thus reducing the number of calculations involved for obtaining the new parameters by *100/(2k-1)* %. In fact, the gain is less because in each iteration the cost for producing the weights w_{ij} dominates. However, it is expected that the gain depends on the maximum observed value since this value determines the number of summands in the calculation of the new estimates. The larger the value of θ_j 's the larger the gain and the larger the sample size the larger the gain.

In order to examine the gain, a small simulation comparison was carried out. For k=2, 100 samples of given a size were simulated for each distribution with parameter vectors *(p ,1,* θ_2 *)*. The time required for ML estimation was calculated for both the general EM algorithm given in Hasselblad (1969) and the improved EM algorithm discussed above. The entries of Table 1, are the relative times namely the ratios of times required by the improved EM algorithm divided by the corresponding times required by the standard EM algorithm. Clearly, almost 20% of the computing time can be saved for k=2.

Table 1

Times for the improved EM algorithm relative to the standard EM algorithm for k=2

p_1	0.25			0.5			0.75		
θ_2	2	5	10	2	5	10	2	5	10
n									
50	82.2	80.3	79.7	82.1	80.0	79.9	83.1	81.5	80.6
100	82.3	80.0	79.2	81.2	79.9	79.7	78.9	80.8	80.0
250	80.9	79.5	79.2	81.3	79.6	79.5	81.9	80.3	79.4
500	81.3	79.3	79.1	81.2	75.2	79.4	80.9	82.0	79.3

Table 2 contains the results for k=3. The vectors of parameters were *(p_1 ,0.3,1,2,* θ_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 the times of the improved EM algorithm, divided by the corresponding times of the standard EM algorithm. Clearly the gain is near 15%.

Table 2
Times for the improved EM algorithm relative to the standard EM algorithm for k=3

p_1	0.25			0.5			0.75		
θ_3	3	5	10	3	5	10	3	5	10
n									
50	86.9	85.9	85.3	86.8	86.3	85.7	87.5	87.0	86.6
100	86.6	86.0	85.3	86.6	86.1	85.4	87.1	86.2	85.8
250	86.3	85.7	85.1	86.2	85.7	85.1	86.3	85.9	85.4
500	86.2	85.6	85.0	85.9	85.5	85.0	86.3	85.8	85.1

The above findings can be useful for ML estimation when the number of support points is not known a priori, see Bohning (1995). Such algorithms usually add one new support point at each step, and try to determine the probability to assign to this point. This procedure usually requires specific numerical methods. However, using the above findings this probability can be determined by simply solving a simple linear equation given by the first moment equation. Further, since a solution with k supports points ought to satisfy the first moment equation, and this is true for the solution with k+1 support points, it is not possible to proceed by simply finding one new support point and assigning to this point some probability. The reason is that since the first moment equation must hold, the algorithm will reject the new point. A combined algorithm which adds a new point followed by some EM iterations for recalculating the probabilities is a preferable procedure.

Behboodian (1969) derived the above scheme independently of Haselblad (1969), for normal mixtures. He also showed that the ML estimates of finite normal mixtures behave in a similar manner.

3. Conclusions

It has been noted that in finite Poisson mixtures one of the ML equations can be replaced by the first moment equation. This simplifies the procedure for ML estimation and can save a lot of computational time. The above results can be generalized for other distributions in the one parameter exponential family, like the exponential distribution. Another very useful result is that the above scheme can be generalized for minimum distance estimation in finite mixtures.

Acknowledgments

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. (1969). 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.

Furman, 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. Willey New York, 2nd edition.

Karlis, D. and Xekalaki, E. (1998). Minimum Hellinger distance Estimation for finite Poisson mixtures. (*To appear in Computational Statistics and Data Analysis).*

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

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

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