Fitting a Normal Mixture to Grouped Fish Length Data

Albert, Jose Ramon, Elloso, Lilia, and Tan, Maria Olivia
Statistical Research & Training Center, Research Division
J&S Bldg., 104 Kalayaan Avenue, Diliman
Quezon City, Philippines
srtcres@srtc.gov.ph

1. Introduction

Empirical data are often either discrete or discretized. Grouping and coarseness of data may be brought about by an attempt to preserve confidentiality, or perhaps, a result of a data gatherer's oblivious selection of a level of accuracy of measurement. For instance, when measuring the length of fish, investigators may find it convenient to record only the frequencies of lengths falling in certain intervals. Grouped data models are examples of incomplete data models. Estimates of the parameters of the assumed underlying distribution may be readily obtained through the EM algorithm (Dempster, Laird and Rubin, 1978) or some variant of this iterative computing scheme (Mclachlan and Krishnan, 1997). Here, we discuss the use of the EM algorithm for real data pertaining to the pooled length distribution of the fish species *auxis thazard* collected in Camotes Sea in July 15th from 1983 to 1987, provided by the Bureau of Fisheries and Aquatic Resources. Here, the component normal distributions may possibly pertain to the length distribution of fish of varying age and/or sex groups

2. Grouped Data from a Finite Mixture

When the log likelihood L_0 of our data is difficult to maximize, we may find a way to augment our data to form some "complete" data set (whose log likelihood L is "easy" to maximize). In this case, we can view the observed data as an incomplete version of the "complete" data set. Since we do not have the log likelihood L (of the complete data) available, we reconstruct it by averaging this log likelihood conditional on the data and some preliminary estimates of the underlying parameters. This forms the E-step. This step essentially provides for an imputation of the missing data components and enables us to perform an M step, i.e. to calculate a pseudo MLE. Repeated iteration of the E step and the M step forms the EM algorithm. For some cases, the EM algorithm may be straightforward to implement, say, for grouped data from a normal distribution, but its numerical convergence may be rather slow. For a finite normal mixture, the EM algorithm not only converges slowly but also has a tendency to get trapped in local maxima of the likelihood function. When these two levels of incompleteness are combined, implementation is not quite straightforward. A simulation experiment reveals that implementing a stochastic version of the EM algorithm rather than the standard EM algorithm generally yields faster numerical convergence.

Note that, in addition to estimating the parameters of the underlying normal mixture model, the number k of normal components also have to be estimated. To go about this problem, we firstly fix the value of k as some small value; estimate the parameters of the k component normal mixture; increase the value of k and repeat the estimation until an "optimal" value of k is chosen (analogous to the forward selection approach in multiple regression modeling). To measure the goodness of fit, a penalty term has to be added to the log likelihood to discourage overparameterization.

Table 1 lists estimates of our parameters together with the values of our model selection criterion MSC=loglikelihood +C*number of independent parameters for varying number of components in the normal mixture model. Here, it is assumed the variance of the different components is constant. Empirical evidence appears to suggest a 3-component normal distribution for the fish length data.

Table 1. Estimates for parameters of fish length mixture distribution

k	Mixing Weights	Estimates of Means	Estimates of Variance	MSC	
				C=2	C=1
2	0.6687898	24.6021004	9.1128996	715.2933	711.2933
	0.3312102	32.3499278			
3	0.3375796	21.89335	1.909613	687.3887	681.3887
	0.477707	27.79028			
	0.1847134	34.77076			
4	0.3248408	21.71781	2.034998	688.3231	680.3231
	0.3757962	27.12237			
	0.1146497	29.31176			
	0.1847134	34.75051			
5	0.3439490	21.92367	2.136815	693.7431	683.7431
	0.0636943	26.64021			
	0.4076433	27.96713			
	0.1592357	34.73977			
	0.0254777	34.95222			

Figure 1 illustrates estimates of the probability density function for the fish length data. These estimates were obtained from the maximum likelihood estimates for varying k. Consistent with the results on the MSC listed in Table 1, there does not appear to be any improvement with the use of more than 3 components for the fish length data.

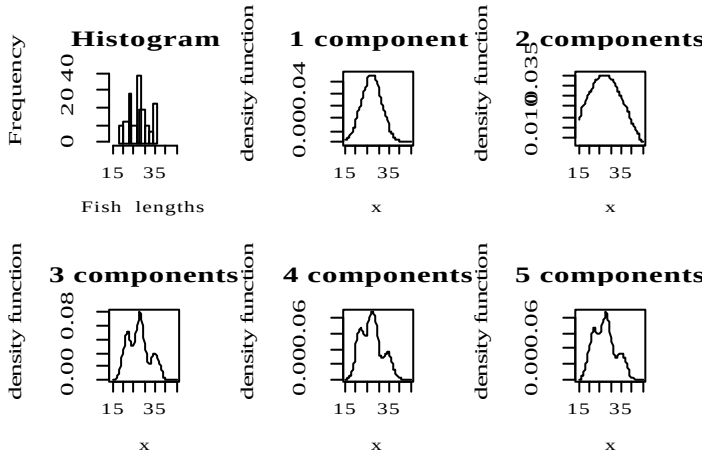

Figure 1. Histogram of fish lengths and probability density function estimates for varying k, k=1,2,3,4,5.

REFERENCES

Dempster, A. P. Laird, N. M. and Rubin, D. B. (1978). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal. Statistical Society. B*, 39, 1-38.

Mclachlan, G. J. and Krishnan, T. (1997). The EM Algorithm and Extensions. New York: John Wiley.

RESUME