Length Frequency Analysis: Maximum Likelihood Estimation of Finite Mixture Model via EM Algorithm and Bayesian Approach

Tony S. H. CHUNG
Census and Statistics Department, Hong Kong, China
tonyshchung@yahoo.com.hk

Abstract

Statistical estimation of a finite mixture model is presented for single length frequency data with the peaks lying on the Von Bertalanffy growth curve. By assuming the asymptotic maximum body size to have a normally distributed prior, we show that the variance formulation proposed by Schnute and Fournier (1980) is a special case of our proposed model. The estimation method utilizes the pseudo-alternating EM algorithm and Newton-Raphson method, while the traditional approach cannot guarantee a Generalised EM sequence if some mixing parameters are not well away from zero. Length frequency data from Northern Abalone is presented to illustrate the method.

Keywords: Finite mixture model, EM algorithm, Bayesian approach

1. Introduction

Beginning with the work of Petersen (1892), fisheries scientists have been attempting to use length frequencies to model growth pattern. The main concept is that length frequencies roughly have modes, at least for juvenile fish, that presumably represent birth cohorts. Lengths at these modes are taken to be the mean lengths for the corresponding cohorts across time, and these mean lengths can then be used for fitting different kinds of growth curves. This paper presents how to build up a statistical model properly by using finite normal mixture distribution in which the means and standard deviations of consecutive components across time obey the Von Bertalanfy growth function that is commonly used by fisheries scientists. The asymptotic maximum body size, L_∞, is assumed to follow a prior distribution (i.e. normal distribution). This Bayesian approach eventually provides us with a more generalised variance formation for the mixture model.

2. Basic Formulation

For the single frequency sample, without loss of generality, it is assumed that only one birth cohort appears in each cycle and the sample is randomly drawn at s time units after the last birth cohort. By Bayes Theorem, for an individual with y unit length and t time units old, we then obtain

$$f(t|y,s)f(y|s) = f(y|t,s)f(t|s). \tag{1}$$

However, the relationship between length and age does not depend on the capture time point. Thus, the conditional density function of y given t remains unchanged no matter when the sample is drawn. Integrating out the age t from both sides, we have a mixture density

$$f(y|s) = \int_{t\in\Psi} f(y|t)f(t|s)dt. \tag{2}$$

The marginal length density depends on the capture time point. For simplicity, it is assumed that $f(t|s)$ is a probability mass function; that is

$$\pi(t) = \Pr(T = t|s). \tag{3}$$

If each cycle contains r time units, given the limited life span of a specified organism, we have

$$f(y) = \sum_{j=1}^{J} \pi_j f(y|t_j) \tag{4}$$

where $t_j = j + \frac{s}{r} - 1$. Let y_i be the class mid-point of ith length interval with h width. Then, assuming $0 < y_1 < \ldots < y_I < \infty$, we define the following notation:

$$\begin{aligned} P(y_i) &= \int_{y_i-h/2}^{y_i+h/2} f(u)du \\ &= \sum_{j=1}^{J} \pi_j \{F(y_i + \frac{h}{2}|t_j) - F(y_i - \frac{h}{2}|t_j)\} \\ &= \sum_{j=1}^{J} \pi_j H(y_i|t_j) \end{aligned} \tag{5}$$

where $F(\cdot)$ is the distribution function of the conditional density function of y given t (i.e. $f(y|t)$) and h is known in advance. Assume $f(y)$ depends on a set of unknown parameters $\theta \in \Theta$. We can obtain the log-likelihood as follows:

$$l(\theta) = \sum_{i=1}^{I} n_i \log P(y_i|\theta). \tag{6}$$

3. EM Algorithm for Mixture Model with Grouped Data

The EM algorithm (Dempster, Laird and Rubin, 1977) is a common method for estimation of a mixture model. We now split the unknown vector parameter θ into two types of parameters (i.e. $\theta = (\phi, \pi)$), where ϕ is the vector parameter of $f(y|t)$ and π is the mixing parameter, $(\pi_1, \cdots, \pi_{J-1})^\top$. Assume we have fully categorized data and let n_{ij} be the number of observations for the ith length group and jth age group, where $\sum_j n_{ij} = n_i$. The likelihood function given all n_{ij}'s is

$$L(\phi|\mathbf{y}, \mathbf{n}) = \prod_{i}^{I} \prod_{j}^{J} H(y_i|t_j, \phi)^{n_{ij}}. \tag{7}$$

Without the age information, the probability density of $n_{i1}, n_{i2}, \cdots, n_{iJ}$ given θ is multinomially distributed as follows:

$$n_i! \prod_{j}^{J} \frac{\pi_j^{n_{ij}}}{n_{ij}!} \tag{8}$$

for $i = 1, \cdots, I$, subject to $\sum_j \pi_j = 1$ and $\sum_j n_{ij} = n_i$. Combining (7) and (8), the log-likelihood (refer to "complete-likelihood" hereafter) could be written as

$$\log L(\theta|\mathbf{n}, \mathbf{y}) = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} [\log H(y_i|t_j, \phi) + \log \pi_j]. \tag{9}$$

The E-step of the EM algorithm requires us to calculate the expected value of the complete-likelihood over the conditional distribution of the missing data, n_{ij}'s, given the observed data, y_i's and n_i's, and current estimates of θ. Given the current estimates at parameter values $\theta = \theta^{(0)}$, the resulting log-likelihood (refer to "pseudo-likelihood" hereafter) is

$$Q(\theta|\theta^{(0)}) = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}^{(0)} [\log H(y_i|t_j, \phi) + \log \pi_j] \tag{10}$$

where, based on (10) and Bayes' Theorem,

$$n_{ij}^{(0)} = E(n_{ij}|y_i, n_i, \theta^{(0)}) \tag{11}$$

$$= n_i \frac{\pi_j^{(0)} H(y_i|t_j, \phi^{(0)})}{\sum_{j=1}^{J} \pi_j^{(0)} H(y_i|t_j, \phi^{(0)})} \tag{12}$$

Obviously, the maximization over π is

$$\pi_{j,EM}^{(1)} = \sum_{i=1}^{I} n_{ij}^{(0)}/n \tag{13}$$

where $j = 1, \cdots, J$, $\sum_j \pi_j = 1$ and $\sum_i n_i = n$. Repeated EM steps maximize the likelihood over θ.

4. Some Extensions of EM Algorithm

Considering the original EM algorithm, if $\pi_{j,EM}^{(p)} = 0$ for some j at step p, then $\pi_{j,EM}^{(p+q)} = 0$ for $q = 0, 1, \cdots$, indicating the sensitivity of the EM algorithm to underspecification of the model. A one-step maximization is too crude to ensure a Generalized EM (GEM) sequence. Even if only one $\pi_{j,EM}^{(p)}$ falls on the zero boundary at the very beginning of the iterations, it will adversely reduce the freedom of the pseudo-likelihood function (i.e. degeneration of the parameter space) and, perhaps, the MLE of such π_j may not lie on the zero boundary. J.R.G. Albert and L.A. Baxter (1995) introduced a *pseudo-alternating EM (PAEM) algorithm* as follows:

1. Decompose $\theta = (\phi, \pi)$ of the generalized parameter vector.
2. Determine a value of ϕ, $\phi^{(p+1)}$ say, such that

   $$Q((\phi^{(p+1)}, \pi^{(p)})|(\phi^{(p)}, \pi^{(p)})) \geq Q((\phi, \pi^{(p)})|(\phi^{(p)}, \pi^{(p)})) \tag{14}$$

   for all ϕ.
3. Determine a value of π, say $\pi^{(p+1)}$, such that

   $$l((\phi^{(p+1)}, \pi^{(p+1)})) \geq l((\phi^{(p+1)}, \pi^{(p)})). \tag{15}$$

The PAEM algorithm can guarantee a non-decreasing sequence of $l(\theta^{(p)})$. If MLE of θ exists, the *PAEM sequence* can reach a stationary point of $l(\theta^*)$ which might be considered as a local maximum. Then, we can obtain the estimates of π_j's from the original-likelihood directly. Since Equation (13) is derived from the first derivative of the pseudo-likelihood, to be more accurate and informative, the Newton-Raphson (NR) method is suggested which involves the second derivative of the original-likelihood function, and the iterative $(p+1)^{th}$ step is defined by

$$\pi^{(m+1)} = \pi^{(m)} - \eta \left[\frac{\partial^2 l(\phi^{(p+1)}, \pi)}{\partial \pi \partial \pi^\top} \right]_{\pi=\pi^{(m)}}^{-1} \frac{\partial l(\phi^{(p+1)}, \pi)}{\partial \pi} \Big|_{\pi=\pi^{(m)}}, \tag{16}$$

with an initial guess $\pi = \pi^{(p)}$ and then $\pi^{(p+1)}$ is obtained while the above equation converges. Fortunately, the first and second derivatives can be derived easily for the mixture model with grouped data. They are

$$\begin{aligned} \frac{\partial l(\phi, \pi)}{\partial \pi_j} &= \sum_{i=1}^{I} n_i \frac{H(y_i|t_j, \phi) - H(y_i|t_J, \phi)}{P(y_i|\theta)} \\ \frac{\partial^2 l(\phi, \pi)}{\partial \pi_j \partial \pi_k} &= -\sum_{i=1}^{I} n_i \frac{[H(y_i|t_j, \phi) - H(y_i|t_J, \phi)][H(y_i|t_k, \phi) - H(y_i|t_J, \phi)]}{P(y_i|\theta)^2}. \end{aligned}$$

5. Von Bertalanffy Growth Function

In this section, we assume the Von Bertalanffy growth function, that is

$$y = L_\infty [1 - e^{-K(t-t_0)}] \tag{17}$$

where y is length (or body size) at age t, K is the growth rate parameter and L_∞ is the asymptotic maximum body size. In practice, there will be measurement error and L may vary among individuals so a reasonable assumption is that:

$$y = L[1 - e^{-K(t-t_0)}] + \epsilon \tag{18}$$

where $\epsilon \sim N(0, \sigma^2)$ and $L \sim N(\mu_L, \sigma_L^2)$,then

$$Y|t \sim N\left(\mu_L[1 - e^{-K(t-t_0)}], [1 - e^{-K(t-t_0)}]^2\sigma_L^2 + \sigma^2\right). \tag{19}$$

Hence, the probability of a fish selected which belongs to ith length interval is

$$P(y_i) = \sum_{j=1}^{J} \pi_j \left\{ \Phi\left(\frac{y_i + \frac{h}{2} - \mu(j + \frac{s}{r} - 1)}{\sigma(j + \frac{s}{r} - 1)} \right) - \Phi\left(\frac{y_i - \frac{h}{2} - \mu(j + \frac{s}{r} - 1)}{\sigma(j + \frac{s}{r} - 1)} \right) \right\} \tag{20}$$

where $\Phi(\cdot)$ is standard normal distribution function and

$$\begin{aligned} \mu(t) &= \mu_L[1 - e^{-K(t-t_0)}] \\ \sigma^2(t) &= [1 - e^{-K(t-t_0)}]^2\sigma_L^2 + \sigma^2 \end{aligned} \tag{21}$$

Then $\phi = (\mu_L, \sigma_L, \sigma, t_0, K)$. Schnute and Fournier (1980) opined that the significance of L_∞, K and t_0 is deceptive. L_∞ may lie far beyond the observed range which could not be verified. It is also difficult to interpret the biological meaning of t_0. Schnute and Fournier recommended reparameterization of (L_∞, K, t_0) to (μ_1, μ_J, k). We modify their recommendation according to additional assumption of normally distributed disturbance terms and prior distribution for L_∞. Let u_j and σ_j be the mean length and the standard deviation of component j. The reparameterization is transforming ϕ into $\phi' = (\mu_1, \sigma_1, \mu_J, \sigma_J, k)$ where

$$\begin{aligned} \mu_1 &= \mu_L[1 - e^{-K(t_1-t_0)}] \\ \mu_J &= \mu_L[1 - e^{-K(t_J-t_0)}] \\ \sigma_1 &= \sqrt{[1 - e^{-K(t_1-t_0)}]^2\sigma_L^2 + \sigma^2} \\ \sigma_J &= \sqrt{[1 - e^{-K(t_J-t_0)}]^2\sigma_L^2 + \sigma^2} \\ k &= e^{-K} \end{aligned} \tag{22}$$

Obviously, the mean length and standard deviation of component j are

$$\mu_j = \mu_1 + (\mu_J - \mu_1)\frac{1 - k^{j-1}}{1 - k^{J-1}} \tag{23}$$

$$\sigma_j = \sqrt{\frac{\sigma_J^2(\mu_j^2 - \mu_1^2) + \sigma_1^2(\mu_J^2 - \mu_j^2)}{(\mu_J^2 - \mu_1^2)}} \tag{24}$$

where $\frac{\mu_J}{\mu_1} \geq \frac{\sigma_J}{\sigma_1}$ and $0 \leq \sigma_1 \leq \sigma_J$. Let us compare the above variance formation to other assumptions used by most fisheries scientists. For example, the standard deviations might be a linear function of the means; that is

$$\sigma_j = \sigma_1 + (\sigma_J - \sigma_1)\frac{\mu_j - \mu_1}{\mu_J - \mu_1} \tag{25}$$

Or, the σ's might be a linear function of ages; that is

$$\sigma_j = \sigma_1 + (\sigma_J - \sigma_1)\frac{j - 1}{J - 1} \tag{26}$$

Schnute and Fournier (1980) combined the above two assumptions and proposed

$$\sigma_j = \sigma_1 + (\sigma_J - \sigma_1)\frac{1 - k^{j-1}}{1 - k^{J-1}} \tag{27}$$

Table 1. Summary Statistics for Goodness of Fit

No.	Age Classes	DF	χ^2	p-value	Likelihood Ratio Statistics	AIC	BIC
1	10	40	43.65	0.319	46.55		
2	7	42	43.33	0.414	46.25		
3	5	52	50.96	0.407	53.75	-1672.58	**-1682.15**
4	6	51	50.02	0.465	51.22	-1672.32	-1682.96
5	7	50	48.87	0.519	49.78	-1672.60	-1684.29
6	8	49	47.01	0.554	48.13	-1672.78	-1685.54
7	7	48	40.48	0.771	42.28	**-1670.85**	-1684.68
8	8	47	40.35	0.743	42.28	-1671.85	-1686.74

DF and χ^2 statistics of models No. 1 and 2 were calculated based on lumping data (i.e. observed frequency less than 1). No. 1 and 2 are quoted from Example 2.3 and 2.4 of Schnute and Fournier (1980). As Schnute and Fournier didn't provide loglikelihood values for No. 1 and 2, the corresponding AIC and BIC values are not available. For model 2, 7 and 8, the last age component of each model either obeys the Von Bertalanfy Growth Assumption nor linear on ages.

No matter what kinds of variance formulation they use, their objective is modelling the increasing trend of σ_j with age.

Equation (21) is derived by relaxing the assumption of a fixed parameter of asymptotic length and explains why the variance increases with age. Furthermore, it also implies, although the same species share the same Brody growth coefficient, k, their lengths for the same age cohort still vary because of the deviation of asymptotic length. If we set $\sigma_L = 0$, then it turns to a finite mixture model with constant variance. On the other hand, if we set $\sigma = 0$, then we get the following relation:

$$\frac{\sigma_1}{\mu_1} = \frac{\sigma_2}{\mu_2} = \ldots = \frac{\sigma_J}{\mu_J} = \frac{\sigma_L}{\mu_L} \tag{28}$$

and Equation (24) will degenerate to exactly the same as that Schnute and Fournier (1980) proposed in (27) except that our parameter space is smaller (i.e $\phi' = (\mu_1, \sigma_1, \mu_J, k)$ since $\sigma_J = \mu_J \frac{\sigma_1}{\mu_1}$). Under the assumption of VBGF and normal distributed asymptotic length, Schnute and Fournier's variance formulation is a special case of our proposed model.

6. Example: Northern Abalone

The dataset considered here pertains to Northern Abalone from Queen Charlotte Islands, British Columbia. Schnute and Fournier (1980) fitted the abalone data by using VBGF type assumption on mean lengths and standard deviations. We will compare their results with ours.

We have found that, if the number of components, J, is less than or equal to four, the Brody growth coefficient, k, converges at the boundary one. It implies the means of all components are equally spaced so that the VBGF becomes invalid. We then start from five components and seek for the most appropriate number of J by the penalized likelihood (i.e. AIC or BIC). The results are summarized in Table 1 and 2. Two well-fitted models from Schnute and Fournier (1980) are also presented (i.e. model 1 and 2). As shown in Table 1, all our suggested models are well-fitted to the data (i.e. p-value > 0.05). AIC yields model 7 while BIC recommends model 3, but the differences among the models are not great. Comparing the likelihood ratio statistics, model 7 and 8 are the best (i.e. better than that of Schnute and Fournier, model 2). Table 2 shows that the asymptotic length is quite sensitive to the Brody growth coefficient, k. But k strongly depends on the number of components. Fewer components imply that the subjects should grow at faster rate to cover the whole range of data. Since there is no other prior information about the number of components, asymptotic length or the growth coefficient except the data itself, we can only say that model 7 is the best fit among all models which has the largest AIC and the smallest likelihood ratio statistic.

To sum up, we agree with Schnute and Fournier that there is still no clear biological conclusions for the abalone data. Breen (1980) obtained a value of $\mu_L = 128.9$ and $k = 0.766$ by using tagging data of abalone (Quayle 1971). The results of

Table 2. Parameter Estimates and Standard Errors

No.	μ_1	μ_J	σ_1	σ_J	k	μ_L	σ_L	σ	t_0	K
1	11.28	111.14	1.84	9.50	0.7916	125.02	-	-	0.5954	0.2337
2	11.29	91.48	1.83	8.06	0.8059	132.78	-	-	0.5882	0.2158
3	11.24	101.88	1.73	13.14	0.9519	518.15	66.67	0.9574	0.5545	0.0492
	(0.391)	(1.195)	(0.341)	(0.920)	(0.024)	-	-	-	-	-
4	11.32	103.71	1.73	12.40	0.8825	210.08	25.01	1.0927	0.5572	0.1250
	(0.386)	(1.197)	(0.339)	(0.996)	(0.020)	-	-	-	-	-
5	11.31	106.08	1.75	11.91	0.8402	157.50	17.59	1.2073	0.5720	0.1741
	(0.385)	(1.204)	(0.343)	(1.015)	(0.016)	-	-	-	-	-
6	11.34	111.44	1.84	10.76	0.8314	149.33	14.28	1.4864	0.5725	0.1847
	(0.391)	(1.199)	(0.373)	(1.140)	(0.012)	-	-	-	-	-
7	11.28	92.40	1.77	9.72	0.8112	136.31	14.21	1.3270	0.5872	0.2093
	(0.390)	(1.028)	(0.353)	(0.860)	(0.016)	-	-	-	-	-
8	11.28	96.50	1.77	10.23	0.7904	123.99	13.03	1.3099	0.5944	0.2351
	(0.389)	(1.098)	(0.350)	(0.991)	(0.016)	-	-	-	-	-

Figures in brackets are standard errors.

model 7 are quite similar to Breen's. To enhance the measure of growth coefficient and number of age groups, multiple length frequency samples and tagging data (i.e. capture-recapture) can highly facilitate these kinds of studies which can include more parametric descriptions of age compositions, birth and mortality rates and transition probabilities.

Acknowledgments

The author wishes to express his appreciation to Dr. John Bacon-Shone for his helpful comments and suggestions on this paper. This paper is an updated, revised, and shorten version of part of the author's unpublished Ph.D. thesis entitled "Statistical Models for Catch-at-length Data with Birth Cohort Information" (Social Sciences Research Centre, The University of Hong Kong, 2005).

References

[1] Aitkin, M. and D.B. Rubin (1985). Estimation and hypothesis testing in finite mixture models. *J. R. Statist. Soc. B.* **47, No.1**, 67-75.

[2] Albert, J.R.G. and L.A. Baxter (1995). Applications of the EM Algorithm to the analysis of life length data, *Appl. Statist.* **44, No. 3**, 323-341.

[3] Breen, P.A. (1980). Measuring fishing intensity and annual production in the abalone fishery of British Columbia. *Canadian Technical Report of Fisheries and Aquatic Sciences.* **No. 947**. Nanaimo, British Columbia.

[4] Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc. B.* **39**, 1-38.

[5] Petersen, C.G.J. (1892). Fiskensbiologiske forhold i Holboek Fjord. *Beret. Dan. Biol. Stn. 1890-1891.* **1**, 121-183.

[6] Quayle, D.B. (1971). Growth, morphometry, and breeding in the British Columbia abalone (*Haliotis kamtschatkana Jonas*). *Fish. Res. Board Can. Tech. Rep.* **No. 279**, 84.

[7] Schnute, J. and D. Fournier (1980). A new approach to length-frequency analysis: growth structure. *Canadian Journal of Fisheries and Aquatic Sciences* **37**, 1337 - 1351.