Possible application of finite normal mixture distribution with a structural model to estimate SBT catch composition from otolith direct aging data

(耳石による直接年齢データからミナミマグロ漁獲組成を推定するための構造モデルと有限正規混合分布の可能な応用)

Hiroshi SHONO and Tomoyuki ITOH (庄野 宏、伊藤智幸)
(National Research Institute of Far Seas Fisheries, Fisheries Research Agency, JAPAN)

Abstract: In this short paper, we introduce a hybrid method using finite normal mixture distribution with a structural model for estimating proportion-at-age of SBT in catches based on the otolith samples and length frequency data. In addition, we discuss the idea for solving some issues subsidiary to the estimation method based on conditional likelihood on the assumption of the use of this hybrid model from the theoretical viewpoint.

要旨: この短い論文では、耳石サンプル及び体長組成データに基づくミナミマグロ漁獲の年齢別割合を推定するための、構造モデルと一緒に有限正規混合分布を使用したハイブリッド法を紹介する。加えて、条件付き尤度に基づく推定法に付随する問題点を解決するための考え方について、このハイブリッドモデルの利用を想定して理論的な観点から論じる。

1. Introduction

Morton and Bravington (2003) proposed the statistical approaches for estimating the age profile of SBT in catches based on the idea of conditional likelihood using the correction of otolith samples and length frequency samples instead of age-length key (ALK) and iterated age-length key (IALK: Kimura and Chikuni, 1987). Their models seem to be effective and reasonable because the disadvantages of ALK and IALK without the assumption of the specific probability distribution are improved. However, their models have some issues as described in Section 3. Therefore, we suppose a hybrid method using finite mixture distribution with a structural model to estimate SBT catch composition from otolith direct aging data in Section 2. We also discuss the idea for solving these issues using this hybrid model from the theoretical viewpoint in Section 3 although our method is not directly related to that by Australian scientists. Properties of maximum likelihood estimator in the finite normal mixture distribution are statistically discussed from the various points of view (Mclachlan and Peel, 2001).

2. Hybrid model (using finite normal mixture distribution with a structural model)

In this section, we describe a hybrid method (using finite normal mixture distribution with a structural model) for estimating proportion-at-age from the otolith samples and length frequency data. The concept and procedure are as follow:

At first, we assume the probability density of finite mixture distribution as:

$$f_\theta(s) = \sum_{a=1}^{m} p_a f_{\theta_a}(s \mid a) \qquad (1)$$

where a: age, m: number of age class, s: length class, $f_\theta(s)$: probability density, p_a: mixing proportion at age a, $f_{\theta_a}(s \mid a)$: conditional density of a given s,

$$\theta := (p_1 ... p_a ... p_m, \theta_1 ... \theta_a ... \theta_m) \quad \left(where \sum_{a=1}^{m} p_a = 1 \right) \tag{2}$$

Next, $s|a$ (conditional random variable s given a) is assumed to be normally distributed with mean μ_a and variance σ_a^2. This means that

$$f_{\theta_a}(s|a) = \frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\left\{ -\frac{(s-\mu_a)^2}{2\sigma_a^2} \right\} \tag{3}$$

where $\theta_a := (\mu_a, \sigma_a^2)$ $(a = 1,...,m)$.

1st step: Assuming the "structure" (such as the following growth curve) on mean and variance of each normal distribution, unknown parameters ($L(\infty), t_0, K, c, \delta$ and $p_a (a = 1,...,m)$ in our case) can be estimated by the maximum likelihood method.

$$\mu_a = L(\infty)[1 - \exp\{-K(a - t_0)\}](von\ Bertalanffy), \quad \sigma_a^2 = c\mu_a^\delta \tag{4}$$

2nd step: Estimate θ (vector of unknown parameters) so as to maximize the following penalized likelihood function.

$$\hat{\theta}_\lambda = \max_\theta \{ l_\lambda(\theta) \} \tag{5}$$

$$l_\lambda(\theta) = (1-\lambda) l(\theta) - \lambda D^{(J)}(\theta(\hat{\xi}), \theta) \tag{6}$$

where

$$D^{(J)}(\theta^*, \theta) := \sum_{a=1}^{m} p_a^* \log \frac{p_a^*}{p_a} + \sum_{a=1}^{m} p_a D(\theta_a^*, \theta_a), \ D(\theta_a^*, \theta_a) := \int f_{\theta_a^*}(s|a) \log \frac{f_{\theta_a^*}(s|a)}{f_{\theta_a}(s|a)} ds \tag{7}$$

(D(p,q) shows the so-called "Divergence of Kullback-Leibler" between p and q),

$l(\theta) := \sum_{i=1}^{n} \log\{ f_\theta(s_i) \} = \sum_{i=1}^{n} \log\{ \sum_{a=1}^{m} p_a f_{\theta_a}(s|a) \}$ (where n: number of observations)

$\theta(\hat{\xi})$: maximum likelihood estimates of the structural model in above 1st step.

3rd step: Estimate the so-called "tuning parameter", λ, using the cross-validation.

$$\hat{\lambda} = \min_{0 \le \lambda \le 1} ACV(\lambda, m) \tag{8}$$

This explicit/analytical calculation for estimating the tuning parameter is very difficult. We can practically compute the approximate cross-validation (ACV) using the concept of generalized information criterion (GIC) proposed by Konishi and Kitagawa (1996).

4th step: Estimate the number of length class (i.e. components), s, in the final step.

$$\hat{m} = \min_m ACV(\hat{\lambda}, m) \tag{9}$$

Detail description and mathematical background of this hybrid method is shown in Eguchi and Yoshioka (2001). Main concept of this model is to integrate the typical finite normal mixture model without any constraints into the adequate "structure" (e.g. von Bertalanffy growth curve and variance function by power exponent) and this objective is to obtain the robust estimators of mixing parameters with reducing the number of parameters. The hybrid model has an additional advantage to estimate the number of components (i.e. number of age-class) shown in equation (9). It is generally difficult to estimate the number of components that it is fixed in many cases.

In case of applying the hybrid model to otolish direct aging samples and length frequency data, it is effective to utilize the estimation of “structure” (i.e. unknown parameters of growth curve). Therefore, in this hybrid model, the otolith samples and length frequency data are corresponding to the parameter estimation of structural model with growth curve and that of full model without any constraint, respectively.

3. Discussions

In this section, we describe several issues of the estimation of SBT catch composition from otolith samples and the method using ALK/IALK or conditional likelihood method by Morton and Bravington (2003). We also discuss the hints/ideas for solving these issues on the assumption of the use of our hybrid model from the theoretical viewpoint.

- Baysian framework in conditional probability distribution of a (age) given s (size)

In the ALK/IALK, the following Bayesian framework (Eqn.(10)) are used.

$$f(a \mid s) = \frac{f(a \cap s)}{f_\theta(s)} = \frac{p_a f_{\theta_a}(s \mid a)}{\sum_{a=1}^{m} p_a f_{\theta_a}(s \mid a)} \tag{10}$$

Although this equation is effective, it seems to be problematic that the conditional density of s given has no statistical error. Therefore, the use of some maximum likelihood function is necessary.

- Assumption of normal distribution of $s \mid a$ (conditional random variable s given a)

Related to above issues, $s \mid a$ is assumed to be normally distributed with mean μ_a and variance σ_a^2 in the conditional likelihood method and hybrid model. This assumption of normality seems to be reasonable judging from the agreed otolith length frequency (submitted by member countries) in each age class.

- Assumption of multinomial distribution of $a \mid s$ (random variable age given size)

In the conditional likelihood method, the assumption of multinomial distribution of a-age given s-size is utilized. This condition seems to be rather strict judging from the agreed otolith age-distribution in each size class and distribution free method (at least regarding proportion-at-age given length-class) such as our hybrid model.

- Estimation of mixing parameters (p_a, μ_a, σ_a^2) (where $a = 1, \ldots, m$ and $\sum_{a=1}^{m} p_a = 1$)

A) Known growth

Because it is not needed to estimate the parameters of normal distribution, $\theta_a = (\mu_a, \sigma_a^2)$, it is a reasonable option to estimate the mixing proportion (p_a) using the conditional likelihood method.

B) Unknown growth

Because the number of unknown parameters seems to be too many, some "structure" (such as growth curve) as used in the hybrid model is available and effective to obtain the robust estimators of mixing parameters (including proportion at age (p_a)).

- Evaluation of the dispersion of unknown parameter (e.g. mixing proportion p_a etc.)

Although it is desirable to use the analytical way by Fisher information as Morton and Bravington (2003) referred, such calculations seems to be very difficult. Therefore, computer intensive method such as parametric bootstrap re-sampling is effective and necessary. Related to this, variance evaluation using ARE (asymptotic relative efficiency) seems not to be good in this case because the number of observations (otolith samples and length frequency data) is not enough to apply the asymptotic theory.

- Similarity of lengh frequency between otolith samples and size composition data

This issue regarding sub-sampling is essential and important. And also degree of similarity may have affected the effective sample size. Therefore, it seems to be necessary to perform various computer simulations changing the length frequency and/or sample size from both otolith samples and size composition data.

There discussions are mainly based on the theoretical statistics. Various computations using agreed otolith data for SBT and simulations using several normal mixture models (including our hybrid method) seems to be necessary for the future.

Acknowledgement

We acknowledge Dr. Sachiko TSUJI, National Research Institute of Far Seas Fisheries, for her useful advice.

References

Eguchi, S. and Yoshioka, K. (2001). Maximum penalized likelihood estimation of finite mixture with a structural model. *Research Memorandom* No.809, 30p. *The Institute of Statistical Mathematics*.

Kimura, D. K. and Chikuni, S. (1987). Mixture of empirical distributions: An iterative application of the Age-Length-Key. *Biometrics*, **43**(1), p.23-35.

Konishi, S. and Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika*, **83**(4), p.875-890.

Mclachlan, G. and Peel, D. (2001). *Finite mixture models*. John Wiley & Sons Inc. 419p.

Morton, R. and Bravington M. (2003). Estimation of age profiles of southern bluefin tuna. CCSBT-ESC/0309/32, 43p.