体長頻度データの年齢組成における有限正規混合分布のハイブリッド成分数推定法

庄野　宏[1†]

Hybrid method for estimating the number of components in a finite mixture of normal distribution from length frequency data to age composition

Hiroshi Shono[1†]

To discuss the problem of estimating age composition from length-frequency data in a finite normal mixture distribution, I introduce a hybrid method for estimating the number of components in the full model without limitation and a structured model with relationship among unknown parameters. The computational procedure was expressed in four steps and statistical interpretation/significance of the hybrid method suggested by Eguchi and Yoshioka in 2001. The hybrid method was applied to the well-known length-frequency data for yellow sea bream, *Dentex hypselosomus*. The computer simulation experiment estimated the number of components in the finite normal mixture model whose probability density function resembles a log-normal distribution. Results showed that the selection performance of the hybrid method for choosing the true model was superior to traditional methods such as Akaike's information criterion（AIC）and/or Baysian information criterion（BIC）.

Key words: age composition, length-frequency data, finite normal mixture distribution, hybrid method, number of components estimation, growth curve

はじめに

様々な魚種の資源量推定において，漁獲物の年齢組成が得られるか否かが手法選択の分岐点になることが多い．年齢組成が得られる場合には，tuned VPA（tuned virtual population analysis）（Gavaris, 1988）などコホート解析手法の利用が可能であるが，そうでない場合には余剰生産量モデル（プロダクションモデル）（Pella and Tomlinson, 1969）を用いることが多い．したがって，tuned VPA の利用に当たっては漁獲物の年齢組成を得るための年齢査定が非常に重要であり，耳石や鱗など生物情報により年齢形質を調べることが魚種を問わず広く行われている．

一方，体長組成データを有限個の確率分布の混合と考え，観測データに基づいて成分数，混合比率および各々の確率密度関数の持つパラメータを推定することも古くから良く行われている．このような計算により生物学的な形質を用いることなく年齢組成の推定が可能となり，コホートに基づく資源量推定が行えるようになる．その場合には，体長組成の年齢分解に正規分布の有限個の混合が利用されることが多い．正規分布のパラメータに関する制約条件を用いないモデル（以後フルモデルと呼ぶ）では，各々の正規分布の混合比率，平均および分散，さらに状況に応じて成分数を体長組成データから推定することができる（Akamine, 1987）．

フルモデルでは柔軟なモデリングが可能だが，最尤推定量の漸近的な性質（一致性・漸近有効性）が成立しないゆえに，モデルの正則性や有界性，識別性等の問題が生じる．実際，成分数を固定した場合には正規分布のパラメータ推定に際して EM アルゴリズムが有効に作用するケースも存在する（赤嶺，2005）．そのため，フルモデルでの成分数推定はかなり難解であり，うまくいかないことも多い．一方，有限混合正規分布においては，正規分布の平均に関して von Bertalanffy の成長曲線などの構造を仮定することが多い（以後構造モデルと呼ぶ）．構造モデルでは，複数パラメータ間の関係式に基づく制約を用いることによりパラ

2011 年 7 月 22 日受付，2012 年 4 月 5 日受理
[1] 鹿児島大学水産学部水産生物・海洋学分野
Fisheries Biology and Oceanography Division, Faculty of Fisheries, Kagoshima University, 4–50–20 Shimoarata, Kagoshima, Kagoshima 890–0056, Japan
[†] shono@fish.kagoshima-u.ac.jp

メータ数が減少し，モデルにおける未知パラメータの推定値が安定するという長所を持つ（Tanaka and Tanaka, 1990; 山川，1997）．

ただし，構造モデルにも短所があり，構造の入れ方，すなわち用いるパラメータ間の関係式に依存するが，一般に構造モデルの成分数推定はフルモデルに比べると難しく，固定して考えることが多い．また，モデルの柔軟性に欠けるため，間違ったモデルを選択するというmis-specificationの恐れがある．例えば，ある成長式よりも別の曲線の方が構造として適当な場合に，mis-specificationによる弊害が生じる．フルモデルと構造モデルの違いは，有限混合正規分布の未知パラメータ（平均・分散および混合比率）間の関係を表す制約式を推定に際して用いるか否かであり，それぞれ上記のような利点と欠点を有する．

そこで，本研究ではフルモデルと構造モデルを融合したハイブリッド法（Eguchi and Yoshioka, 2001）を取り上げ，その計算メカニズムを紹介すると同時に，この方法を広く知られているキダイ（*Dentex hypselosomus*）の体長組成データ（田中，1956）に適用し，有限混合正規分布のパラメータおよび成分数の推定を行った．併せて，体長のモードや各成分の識別が難しい有限混合正規分布モデルを利用した計算機シミュレーションを行い，成分数推定におけるハイブリッド法の優位性を実証する．さらに，Akaike（1973）のAIC（Akaike's information criterion）やSchwarz（1978）のBIC（Bayesian information criterion）などの代表的な情報量規準やフルモデルのパラメータの一部に事前分布を取り入れたBayes型規準（庄野，2006）を用いて，正しい成分数を推定するためのモデル性能の比較を，フルモデルとハイブリッド法による計算機実験に基づいて行った．

材料と方法

フルモデルの成分数推定における情報量規準の利用

有限正規混合モデル（制約なしフルモデル）の定式化は以下のとおり：体長組成データの年齢分解の例では，式（1）におけるXが体長を表す．

$$X_1,\ldots,X_n \sim (i.i.d.)\quad p.d.f.\, g(x|\theta)=\sum_{j=1}^{m}\alpha_j f(x|\mu_j,\sigma_j^2) \quad (1)$$

ただし，X（標本ベクトル），$\theta=(\mu_1,\ldots,\mu_m;\ \sigma_1^2,\ldots,\sigma_m^2;\ \alpha_1,\ldots,\alpha_{m-1})$（未知パラメータベクトル），$m$（成分数），$\mu$（正規分布の平均），$\sigma^2$（正規分布の分散），

$$f(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

（正規分布の密度関数），

$$\sum_{j=1}^{m}\alpha_j=1,\quad \alpha_j>0\ \ (j=1,\cdots,m)$$

とする．

このモデルの対数尤度関数は以下のようになる．

$$l(\theta|X)=\sum_{i=1}^{n}\log\sum_{j=1}^{m}\alpha_j f(x_i|\mu_j,\sigma_j^2)$$
$$=\sum_{i=1}^{n}\log\sum_{j=1}^{m}\alpha_1\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right\} \quad (2)$$

ただし，nは標本数とする．

Leroux（1992）は，対数尤度関数にペナルティ項を付け加えた罰金付き最尤法（penalized maximum likelihood method）による推定法を提案している．これは，対数尤度関数の（−2）倍にペナルティ項を付け加えた式：

$$P(\theta,m)=-2\log L(\theta|X)+2a_{m,n}=-2l(\theta|X)+2a_{m,n} \quad (3)$$

（ただし，$a_{m,n}$は$a_{m,n}>0,\ a_{m+1,n}>a_{m,n}$，$(a_{m,n}/n)\to 0\ (n\to\infty)$を満たす実数列とし，$L$は尤度関数を表す）により，成分数$m$の推定量$\hat{m}$を最小化問題$\min_m\{\min_\theta P\}$の解として求める方法である．

（3）式のペナルティ項の設定により，複数の情報量規準が表現できる．

$$a_{m,n}=\left\{\begin{array}{l} 3m-1 \quad (\mathrm{AIC}) \\ \dfrac{(3m-1)\log(n)}{2} \quad (\mathrm{BIC}) \end{array}\right\} \quad (4)$$

次に，（3）式の変形により具体的に表現されるBayes型モデルについて記述する．$a_{m,n}=\log p(\alpha)$（ただし$\alpha=(\alpha_1,\ldots,\alpha_m)$）とおくことにより，有限混合正規分布の混合比率$\alpha$に対する事前分布$p_{m,n}(\alpha)$を取り入れることが可能となる．また，他のパラメータの扱いは最尤推定の場合と同様である．実際，αの事前分布に対応する事後分布は，有限混合正規分布の尤度関数と事前分布の積，すなわち下式（5）で表される．なお，このBayes型モデルは（4）式のBICとは異なることに注意が必要である．

$$P(\theta|X)=\left\{\prod_{i=1}^{n}\left[\sum_{j=1}^{m}\alpha_j f(x_i|\mu_j,\sigma_j^2)\right]\right\}p_{m,n}(\alpha) \quad (5)$$

Leroux（1992）は，論文中でこの罰金付き最尤法による成分数の推定量$\hat{m}$について言及し，この値が漸近的に過小推定にならないことを示している．しかし，漸近的な過大推定の可能性については言及していないため，漸近的な一致性は証明されておらず（後述のハイブリッド法においても証明はなされていない），適用に当たっては注意が必要である．一般に，有限混合分布の漸近的な一致性は，仮に理論的に成り立つ場合においても標本数を大きくした場合に真の値に近づく収束のスピードが極めて遅いと考えられ

ており，標本数が数百ないし数千程度の場合には，実用上意味を持たない可能性もある．なお，体長組成の年齢分解など, 水産資源分野においてフルモデルの成分数推定でカイ二乗検定を用いる事例もいくつか見受けられるが，情報量規準の場合とは異なり，最尤推定量の漸近的な性質が成立しないため，統計学的に見れば不適切である（付録参照）．

Leroux（1992）の罰金付き最尤法に類似する成分数推定法として，最小距離法（minimum distance method; Henna, 1985）や罰金付き最小距離法（penalized minimum distance; Chen and Kalbfleisch, 1996）などが挙げられるが，罰金付き最尤法に比べて特に優れているとは言えないため，本稿では省略する．これらの手法も含めた有限混合分布における成分数やパラメータ推定法の発展については，McLachlan and Peel（2000）に詳説されている．

有限混合分布におけるパラメータ推定のためのハイブリッド法

Eguchi and Yoshioka（2001）は，フルモデルの成分数推定に関する対数尤度関数の過剰な振る舞いの問題を解決するために，フルモデルと構造モデルの概念を融合させた有限混合モデルにおける新しい成分数，およびパラメータ推定法（以下ハイブリッド法と呼ぶ）を提案した．そこで，本節ではハイブリッド法による計算手順についてEguchi and Yoshioka（2001）に従い，有限正規混合モデルを用いて，4つのstepに分けて簡潔に述べる．

Step-1：構造モデルH: $\theta=\theta(\xi)$の下で，最尤法によりパラメータの推定値$\theta(\hat{\xi})$を求める．ただし，ξは構造モデルのパラメータベクトルを表す．このプロセスは構造モデルにおけるパラメータ推定手順と全く同じである．一例ではあるが，体長組成データの年齢分解においては，正規混合分布の成分数を固定し，平均に対するvon Bertalanffyの成長曲線：

$$\mu(t)=L_\infty\{1-\exp[-K(t-t_0)]\} \tag{6}$$

（ただし，L_∞，K，t_0は最大到達体長，成長係数，体長0となる年齢を表す）や，分散の形状，すなわち平均と分散の関係：

$$\sigma^2(t)=c\{\mu(t)\}^p \tag{7}$$

（ただし，c，pはそれぞれ定数およびべき乗数を表す）を仮定し，パラメータ（L_∞，t_0，K，c，p）の推定を行えば良い．

Step-2：罰則付き尤度関数

$$l_\lambda(\theta)=(1-\lambda)l(\theta)-\lambda D^{(J)}(\theta(\hat{\xi}),\theta) \tag{8}$$

（ただし，チューニングパラメータλは0以上1以下の値を取る）を最大にするパラメータθの値を求める．(8) 式では，
$l(\theta)$：フルモデルの対数尤度関数
$D^{(J)}(\theta(\hat{\xi}),\theta)$：構造モデルとフルモデルのKullback–Leibler divergence（K–L情報量）を表す．成分数mの有限正規混合分布の場合には，

$$D^{(J)}(\theta^*,\theta)=\sum_{j=1}^{m}\alpha_j^*\log\frac{\alpha_j^*}{\alpha_j}+\sum_{j=1}^{m}\alpha_j^* D(\omega_j^*,\omega_j) \tag{9}$$

（ただし$\omega=(\mu,\sigma^2)$とする），

$$\alpha_j f(x,\omega_j)=\alpha_j f(x\,|\,\mu_j,\sigma_j^2)=\frac{\alpha_j}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right\} \tag{10}$$

$$\alpha_j^* f(x,\omega_j^*)=\alpha_j^* f(x\,|\,\mu_j^*,\sigma_j^{2*})=\frac{\alpha_j^*}{\sqrt{2\pi\sigma_j^{2*}}}\exp\left\{-\frac{(x-\mu_j^*)^2}{2\sigma_j^{2*}}\right\} \tag{11}$$

（ただし，fは正規分布の確率密度関数を表す）
となり，(10)，(11) 式はそれぞれフルモデルと構造モデルの密度関数を表す．

Step-3：クロスバリデーション（CV: cross-validation）（または近似的なクロスバリデーション（ACV: approximate cross-validation））により，Step-2における加重平均の割合を定めるチューニングパラメータλの値を求める．

$$\hat{\lambda}=\arg\min_{0\le\lambda\le1} CV(\lambda,m)\ (\text{or}\ \arg\min_{0\le\lambda\le1} ACV(\lambda,m)) \tag{12}$$

データをランダム分割して予測を行うクロスバリデーションに際して，全データから第k番目の観測値$x(k)$ (k=1, …, n)を抜いてのジャックナイフ法に類似する計算は，標本数nや成分数mが大きくなると実行が難しいため，(13) 式の近似的な方法による最小化も用いられている．

$$ACV(\lambda)\approx-\frac{1}{n}\sum_{i=1}^{n}\log f(x_i\,|\,\hat{\theta}_\lambda)+\frac{1-\lambda}{n}\dim(\theta)+\frac{\lambda}{n}\dim(\xi) \tag{13}$$

Step-4：最後に下式（14）のクロスバリデーション，もしくは近似的なクロスバリデーションにより，成分数mの推定を行う．

$$\hat{m}=\arg\min_m CV(\hat{\lambda},m)\ (\text{or}\ \arg\min_m ACV(\hat{\lambda},m)) \tag{14}$$

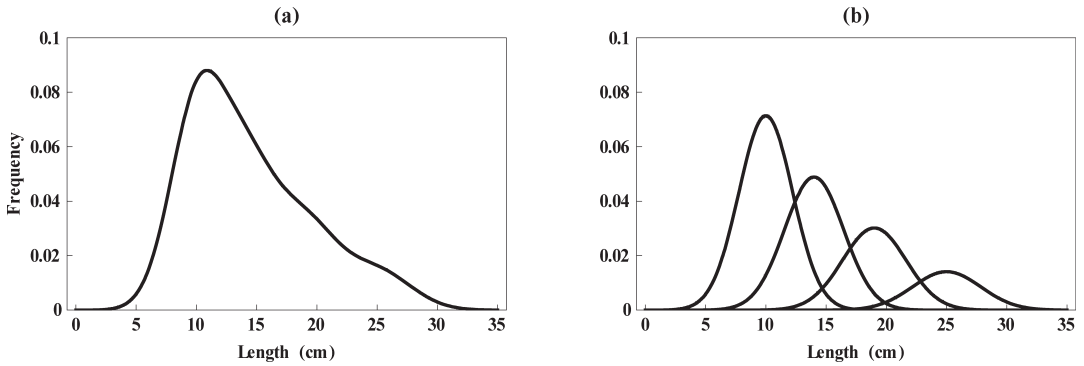

Figure 1. Probability density function of normal finite mixture distribution used for the computer simulation experiment, modified from Shono（2006）.（a）: Plot of 4-component normal mixture densities shown in Eq.（19）,（b）: plot of probability density function in each normal distribution.

Step-4 での近似はGIC（generalized information criterion: Konishi and Kitagawa, 1996）を用いることにより実行可能である．GICは統計的汎関数に基づいて提案された情報量規準であり，有限正規混合分布を含む確率密度関数の場合にはAICを精密化した情報量規準であるTIC（Takeuchi's information criterion: 竹内，1976）とGICとが一致するため，TICで代用可能となる．

漁業データへの適用とシミュレーション実験

最初に，有名なキダイ（*Dentex hypselosomus*）の体長組成データ（田中，1956）を用いて，ハイブリッド法により有限正規混合分布の成分数および各正規分布のパラメータ（平均と分散，混合比率）の推定を行った．

このキダイデータへの過去の適用例では成分数を固定した場合が多く，例えばAkamine（1987）では成分数を5と固定しているが，その是非については議論していない．本論文では以下の構造を仮定して，成分数も含めたパラメータ推定を行った．

$$\mu(t)=\mu(t|L_\infty, K, t_0)=L_\infty[1-\exp[-K(t-t_0)]\} \quad (15)$$

（平均–von Bertalanffy成長式）

$$\sigma^2(t)=\sigma^2(t|c, p)=c\{\mu(t)\}^p \quad (16)$$

（分散–平均のべき乗）

$$\alpha_i(t)=\alpha_i(t\,|\,K)=\exp(-ikt)\Big/\sum_{i=1}^{m}\exp(-ikt) \quad (17)$$

（混合比率–減少率一定の指数関数）
（ただしkは混合比率における減少率の度合いを表すパラメータである）

$$m\in N\text{（成分数）} \quad (18)$$

と仮定し，Eguchi and Yoshiokaのハイブリッド法による解析を行った．

次に，成分数が4であり，一見したところ各々の正規分布の識別が難しく，1つの対数正規分布のように見える（19）式のモデルから乱数を100回発生させて，成分数mが4と正しく推定される回数を計算した．

$$X_1, \ldots, X_n\sim(i.i.d.)g(x)=0.4f(x|10,5)+0.3f(x|14,6)+0.2f(x|19,7)+0.1f(x|25,8) \quad (19)$$

（ただし$f(x|\mu, \sigma^2)$：平均μ，分散σ^2に従う正規分布の密度関数を表し，標本数nは250と設定した）

Figure 1にモデルの確率密度関数および個々の正規分布を重ね書きした確率密度関数を示す．このハイブリッド法の適用に際しては正規分布の平均および分散に関する（15），（16）式の構造を仮定しているが，キダイの例とは異なり（17）式の混合比率に対する制約を使用していない．

結　果

キダイデータへのハイブリッド法の適用では，下記の推定値が得られた．

L_∞=48.7, K=0.13, t_0= − 1.08, c=0.62, p=0.24, m=4,
k=0.55 (α_1=0.41, α_2=0.29, α_3=0.19, α_4=0.11)

これを図示するとFig. 2のようになる．成分数のみならず構造モデルに基づく正規分布のパラメータも含め，全体的な当てはまりが良いことはFig. 2における有限混合正規分布の形状から見てとれる．

また，キダイの体長組成データの成分数推定をAIC，BIC，Dilichlet事前分布を用いたBayes型規準およびMLL

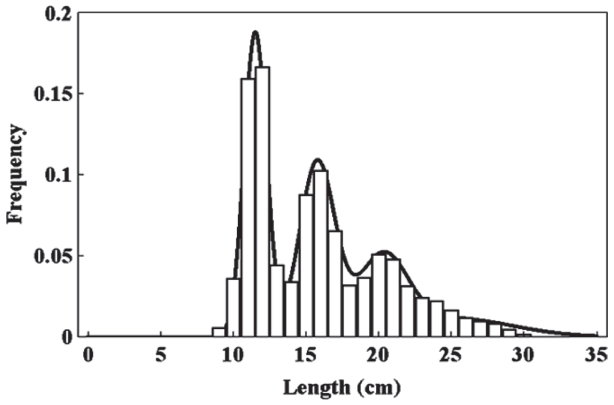

Figure 2. Fitting of the hybrid method to the length-frequency data for yellow sea bream, *Dentex hypselosomus*（Tanaka, 1956）.

Table 1. Values of each information criterion in the length-frequency data for yellow sea bream, *Dentex hypselosomus*（Tanaka, 1956）. Numbers in bold type show the minimum value in each criterion. # of comp. means the number of components.（−2）×MLL: 75342.0, AIC: 75376.0, BIC: 75472.6, Bayes-type: 75371.6.

# of comp.	3	4	5	6
(−2)×MLL	75479.6	75367.6	75351.6	**75342.0**
AIC	75495.6	75389.6	75379.6	**75376.0**
BIC	75556.0	**75472.6**	75485.2	75504.4
Bayes-type	75486.6	75379.6	**75371.6**	75372.2

（maximum log likelihood（最大対数尤度）：計算にはMLLを（−2）倍した数値を使用）という4つの規準に基づいて行った．これらの指標の値はTable 1のようになり，（−2）×MLLとAICでは6，BICでは4，Bayes型規準では5，と使用する情報量規準により推定された成分数が異なる結果が得られた．一方，ハイブリッド法では有限正規混合分布の成分数が4と推定されたため，結果の解釈には注意を要する．

なお，このキダイデータを用いて年齢査定を行った研究（真道，1960; Oki and Tabeta, 1998）における成長式の推定結果は，田中（1956）が成分数を5と固定して算出した各正規分布の平均の推定値とよく符合しており，生物学的な観点から成分数を4とすべきかどうかは議論の余地がある（山田ほか，2007）.

次に，Fig. 1の確率密度関数からの乱数を利用して著者が過去に行ったシミュレーション実験の結果（庄野，2006）を紹介する．ここでは標本数を3000および7000と設定し，MLLの（−2）倍，AIC，BIC，Bayes型という4つの情報量規準により，繰り返し数100の計算機実験において，成分数が4（取り得る範囲は2から6と仮定）と正しく推定される回数を計算した.

Table 2. Result of model selection for MLL（maximum log-likelihood）, AIC, BIC and Bayes-type information criterion based on the computer simulation experiment used in the finite mixture of normal distribution（Shono, 2006）. Bold type shows the maximum number in each criterion. # of comp. means the number of components.

Case-A: sample size-3000

# of comp.	2	3	**4（true）**	5	6
MLL			**66**	1	33
AIC			**90**	9	1
BIC		11	**89**		
Bayes-type		5	**94**	1	

Case-B: sample size-7000

# of comp.	2	3	**4（true）**	5	6
MLL			**93**		7
AIC			**98**	1	1
BIC		2	**98**		
Bayes-type			**100**		

Bayes型規準では，混合比率の事前分布としてDilichlet分布

$$\pi(\alpha) = c\prod_{j=1}^{m}\alpha_j^{-1/2} \qquad (20)$$

（ただし$c=\Gamma(m/2)\pi^{m/2}$, $\alpha=(\alpha_1, \ldots, \alpha_m)$, $\alpha_m=1-\sum_{j=1}^{m-1}\alpha_j$ とおく）

を利用した．これは，Dilichlet分布

$$\pi(\alpha\,|\,\beta) = \frac{\Gamma\left(\sum_{j=1}^{m}\beta_j\right)}{\prod_{j=1}^{m}\Gamma(\beta_j)}\prod_{j=1}^{m}\alpha_j^{\beta_j-1} \qquad (21)$$

（ただし$\beta=(\beta_1, \ldots, \beta_m)$とする）

において$\beta_1=\beta_2=\cdots=\beta_m=1/2$と仮定した無情報事前分布であるが，この超パラメータを階層Bayes法や経験Bayes法などにより推定することも可能である．結果は，Table 2のとおりであるが，Bayes型規準やAIC，BICを用いた真の成分数を選ぶ場合の性能は，標本数が大きい場合には優れている．その中でも特にBayes型規準の選択パフォーマンスが良く，目で見て判別が難しい分布においても良好である（庄野，2006）.

これに対して，ハイブリッド法で正しい成分数を選ぶパフォーマンスが，Bayes型規準と比較して高いという優位性を示すため，標本数を250に設定して同じ100回の繰り返しを持つ計算機実験を行った．その結果，ハイブリッド

Table 3. Result of model selection for hybrid method based on the computer simulation experiment used in the finite mixture of normal distribution. Bold type shows the maximum number in this method. # of comp. means the number of components.

# of comp.	2	3	**4（true）**	5	6
Hybrid-type		2	**97**	1	
Bayes-type		10	**80**	10	

法の真の成分数4を選ぶという選択パフォーマンスは97%となり，Bayes型規準の80%と比べて非常に良くなっている（Table 3）．さらに，情報量規準の中で性能が一番良いBayes型規準において，標本数が3000の場合に真の成分数4を選択する確率が94%であるのに対し，ハイブリッド法では，標本数が250と少ない場合でも正しい成分数4を選ぶ確率が97%と高いことが，実験により示されている．

考　察

前節の結果を総合すると，フルモデルと構造モデルを融合させたハイブリッド法（Eguchi and Yoshioka, 2001）による有限混合分布のパラメータ推定では，理論的のみならず計算機実験においても正しい成分数を選択するパフォーマンスが，AICなどの情報量規準によるそれに比べて高く，実用的である．特に，水産資源における体長組成の年齢分解では，魚類の成長曲線などの利用できる構造に関する情報も多く，正規分布の仮定の妥当性と合わせて，その適用範囲は広いと考えられる．

本論文ではハイブリッド法による成分数推定に焦点を合わせているが，この方法はその他のパラメータ（構造モデルのパラメータや構造モデルを通して推定される各正規分布のパラメータ）の推定にも有効である．実際，ハイブリッド法で推定されたキダイの体長組成に関する有限正規混合分布の密度関数の観測データから得られるヒストグラムに対する当てはまり（Fig. 2）から判断すると，成分数以外の各正規分布に関するパラメータも良く当てはまっているため，成分数と同時に他のパラメータをこの方法で推定することも，実用上の問題は生じないと考えられる．

一方，解析における一番の目標は成分数推定のため，ハイブリッド法で成分数を推定し，その後に残りのパラメータを（成分数を固定した）フルモデルや構造モデルで推定することも合理的と思われる．フルモデルの場合には，成分数推定に際して最尤推定量の漸近的な性質である一致性および漸近正規性が成り立たない，という対数尤度の過剰な振る舞いに起因する問題を避けることが可能である．構造モデルでは，パラメータ推定値の安定性が増加し，より柔軟なモデリングが可能となる．この対数尤度の過剰なふるまいに起因する問題は，成分数を推定する際に生じるため，成分数を固定した場合の構造モデルやフルモデルでは，考慮すべき事項とはなりにくい．

このように，成分数以外のパラメータ推定はハイブリッド法でも構造モデルやフルモデルによる最尤法でも可能である．さらに，式（8）の罰則付き対数関数に着目して，チューニングパラメータλの値が1に近い場合には構造モデルを，0に近い場合にはフルモデルを利用し，それ以外の中間的な値の場合にはハイブリッド法で推定することも現実的と思われる．また，パラメータの標準誤差の評価について，構造モデルやフルモデルに基づく場合にはFisher情報行列の利用が可能であり，ハイブリッド法を利用する場合にはBootstrap法など実験的な手法の適用が妥当と考えられる．

なお，ハイブリッド法や通常の構造モデルでは，仮定した構造のmis-specificationの影響が大きいゆえに弊害が生じる可能性が存在する．例えば，成長式としてGompertz曲線の構造を仮定したが，von Bertalanffy曲線の方が当てはまりが良い場合などである．このようなmis-specificationを防ぐためには，最初に構造を導入するStep-1の段階で，候補となるいくつかの構造を用いて最尤法によるパラメータ推定を行い，尤度関数もしくはAICなどの情報量規準の値を比較することが有益である．さらに，これら複数の候補となる構造の優劣は，チューニングパラメータλおよび成分数mの推定を経て変わることもありうるため，Step-4の終了後に推定されたλとmの値を利用して（Step-2における）罰則付き尤度関数を求め，比較することも有用である．

最後に，この手法は理論的に複雑であり，統計パッケージや数式処理ソフトウェアによる実装も厄介である（特にStep-3およびStep-4）．著者はハイブリッド法の解析にMathematica Ver.7およびVer. 8（Wolfram Research Inc.）を用いたが，標本数が多い場合に多大な時間を要し，計算が途中で止まってしまうことも多く見られた．そのため，さらなる普及のためにはハイブリッド法のアルゴリズム，特にStep-3およびStep-4で用いられる近似的なクロスバリデーション（approximate cross-validation: ACV）の方法論の開発および改良と合わせて，統計パッケージや数式処理ソフトウェアなどを利用した計算過程の自動化，もしくは高級言語によるサブルーチン化が望まれる．

謝　辞

本研究における有限正規混合分布のパラメータ推定，特に成分数推定に関して多くの議論並びに助言をいただいた，有限混合分布における成分数推定法（ハイブリッド法）の提案者である統計数理研究所教授江口真透博士および国士舘大学教授吉岡耕一博士に，深く感謝の意を表する．有益なコメントを下さった査読者の方々に厚くお礼申し上げる．

引用文献

Aitkin, M. and D. B. Rubin (1985) Estimation and hypothesis testing in finite mixture models. J. Roy. Stat. Soc. B Met., **47**, 67–75.

Akaike, H. (1973) Information theory and an extention of the maximum likelihood principle. In. 2nd International Symposium on Information Theory. eds. B. N. Petrov and F. Csaki, Akadémiai Kiadó, Budapest, 267–281.

Akamine, T. (1987) Comparison of algorithms of several methods for estimating parameters of a mixture of normal distributions. Bull. Jap. Sea Reg. Fish. Res. Lab., **37**, 259–277.

赤嶺達郎（2005）混合正規分布とEMアルゴリズム. 水産海洋研究, **69**, 174–183.

Chen, J. and J. D. Kalbfleisch (1996) Penalized minimum-distance estimates in finite mixture models. Can. J. Stat., **24**, 167–175.

Chernoff, H. (1954) On the distribution of the likelihood ratio. Ann. Math. Stat., **25**, 573–578.

Eguchi, S. and K. Yoshioka (2001) Maximum penalized likelihood estimation of finite mixtures with a structural model. Inst. Stat. Math. Mem., **809**, 30 pp.

Gavaris, S. (1988) An adaptive framework for the estimation of population size. CAFSAC Res. Doc., **88/29**, 12 pp.

Hartigan, J. A. (1985) A failure of likelihood asymptotics for normal mixture. In. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. 2, eds., L. M. LeCam and R. A. Olshen, Wadsworth Monterey, 807–810.

初山光司・狩野　裕・長尾壽夫（1995）Selecting the number of components in a mixture of normal distributions: A simple case.（有限正規混合分布における成分数の選定：単純な場合）. 数理解析研究所講究録, **916**, 131–148.

Henna, J. (1985) On estimating of the number of constituents of a finite mixture of continuous distributions. Ann. Inst. Stat. Math., **37**, 235–240.

Konishi, S. and G. Kitagawa (1996) Generalised information criteria in model selection. Biometrika, **83**, 875–890.

Leroux, B. G. (1992) Consistent estimation of a mixing distribution. Ann. Stat., **20**, 1350–1360.

McLachlan, G. J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Appl. Stat., **36**, 318–324.

McLachlan, G. J. and D. Peel (2000) Finite Mixture Models. Wiley Series in Probability and Statistics, Applied probability and statistics section, Wiley-Interscience, New York, 419 pp.

Oki, D. and O. Tabeta (1998) Age, growth and reproductive characteristics of the yellow sea bream, *Dentex tumifrons*, in the East China Sea. Fish. Sci., **64**, 191–197.

Pella, J. J. and P. K. Tomlinson (1969) A generalized stock production model. Bull. IATTC, **13**, 421–496.

Schwarz, G. (1978) Estimating the dimension of a model. Ann. Stat., **6**, 461–464.

Shapiro, A. (1985) Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. Biometrika, **72**, 133–144.

真道重明（1961）東海におけるレンコダイ資源の研究. 西水研研報, **20**, 1–198.

庄野　宏（2006）モデル選択手法の水産資源解析への応用―情報量規準とステップワイズ検定の取り扱い―. 計量生物学, **27**, 55–67.

竹内　啓（1976）情報統計量の分布とモデルの適切さの規準. 数理科学, **153**, 12–18.

Tanaka, E. and S. Tanaka (1990) A method for estimating age-composition from length-frequency by using stochastic growth equation. Nippon Suisan Gakkaishi, **56**, 1209–1218.

田中昌一（1956）Polymodalな度数分布の一つの取扱方及びそのキダイ体長組成解析への応用. 東海区水研研報, **14**, 1–13.

山田梅芳・堀川博史・中坊徹次・時村宗春（2007）東シナ海・黄海の魚類誌. 東海大学出版会, 神奈川, 1262 pp.

山川　卓（1997）複数体長組成データの解析によるイセエビの成長と齢別組成および加入の推定. 水産海洋研究, **61**, 23–32.

付　録

対数尤度の過剰なふるまいについて

―有限混合分布における成分数推定のカイ二乗検定の利用に関する注意―

フルモデルにおける有限混合分布の成分数推定に関し，対数尤度の過剰なふるまい，すなわち最尤推定量の漸近的性質（一致性・漸近有効性）が成り立たないゆえに推定がうまくいかないケースが存在する．例えば，成分数の推定に際してカイ二乗検定を用いることができない．そこで，式（A.1）の簡略化されたモデルを用いて，このことについて説明する．以下，有限混合分布に関する正規分布の仮定は，必ずしも必須ではない．

$$X_1, \ldots, X_n \sim (i.i.d.)p(x|\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \alpha f(x|\mu_1, \sigma_1^2)+(1-\alpha) f(x|\mu_2, \sigma_2^2) \quad \text{（A. 1）}$$

（ただし，$X_1, \ldots, X_n$：標本ベクトル，n：標本数，$f(x|\mu, \sigma)$：平均μ，分散σ^2）に従う正規分布の密度関数，α：2つの正規分布の混合比率とする）において，帰無仮説および対立仮説を以下のように設定する．

$$\left.\begin{matrix} H_0 : \alpha = 1 \\ H_1 : 0 \le \alpha < 1 \end{matrix}\right\} \quad \text{（A. 2）}$$

このときの尤度比検定統計量は，

$$\lambda_n = \frac{\max_{H_0} \prod_{i=1}^{n} p(x \mid \alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)}{\max_{H_0 \cup H_1} \prod_{i=1}^{n} p(x \mid \alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)} \quad \text{（A. 3）}$$

と表される．ここでは最尤推定量の漸近的な性質（一致性および漸近有効性）が成り立たないため，一般に通常の漸近カイ二乗近似

$$-2\log\lambda_n \rightarrow \chi^2_{5-2}(=\chi^2_3) \text{（}n \rightarrow \infty \text{のとき）} \quad \text{（A. 4）}$$

は成立しない．例えば，式（A.5）の場合には$-2\log\lambda_n \rightarrow \infty$（$n \rightarrow \infty$のとき）となり，無限大に発散する（Hartigan, 1985）：

$$X_1, \ldots, X_n \sim (i.i.d.)p(x|\alpha, \mu)=\alpha f(x|0, 1)+(1-\alpha)f(x|\mu, 1) \quad \text{(A.5)}$$

その理由として

(1) 認定可能性の欠如

(2) 帰無仮説の下で真値α=1がパラメータ空間の内点にならない

(3) 帰無仮説の下でFisher情報量が特異

などが挙げられ，特に（2）および（3）は，最尤推定量の漸近的な性質が成立しない原因となっている．

以上の3つの理由により，混合分布モデルにおける成分数やパラメータ数の推定に際して，一般に漸近カイ二乗近似による尤度比検定を使用することができない．しかし，このことは尤度比検定統計量に対して$-2\log\lambda_n$の漸近分布が求まらない，ということを必ずしも意味するものではない．

実際，Shapiro（1985）は，$-2\log\lambda_n$の漸近分布が特別な場合にカイ二乗バー分布と呼ばれるカイ二乗分布の重み付き平均の分布になることを，Chernoff（1954）による関数解析的な手法を用いて証明している．なお，ここでいう特別な場合とは，（A.1）の例で考えると分布に関するパラメータが共通（$\mu_1=\mu_2$ or $\sigma_1^2=\sigma_2^2$），あるいは分布形が完全に既知（$\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$のすべてが既知or nuisance parameter）であるような場合を指している．また，これらの特別の場合におけるカイ二乗バー検定については，いくつかの事例がMcLachlan and Peel（2000），初山ほか（1995）などに記載されている．その他では，McLachlan（1987）が，有限混合正規分布モデルで分散が等しい場合に計算機シミュレーションを用いて，$-2\log\lambda_n$の分布は漸近的に自由度が（混合比率を除いたパラメータ数の差)×2となるカイ二乗分布に従う（式（A.1）の場合は自由度が4となる），と述べている．また，Aitkin and Rubin（1985）は，混合比率による事前分布の導入とは異なるBayes統計学の概念を利用した尤度比検定法を提案している．