Parameterizing probabilities for estimating age-composition distributions for mixture models

Daniel K. Kimura

Martin W. Dorn

Alaska Fisheries Science Center
National Marine Fisheries Service
7600 Sand Point Way N.E.
Seattle, Washington 98115-6349
E-mail (for D. Kimura): Dan.Kimura@noaa.gov

When estimating parameters that constitute a discrete probability distribution $\{p_j\}$, it is difficult to determine how constraints should be made to guarantee that the estimated parameters $\{\hat{p}_j\}$ constitute a probability distribution (i.e., $\hat{p}_j>0$, $\Sigma\hat{p}_j=1$). For age distributions estimated from mixtures of length-at-age distributions, the EM (expectation-maximization) algorithm (Hasselblad, 1966; Hoenig and Heisey, 1987; Kimura and Chikuni, 1987), restricted least squares (Clark, 1981), and weak quasisolutions (Troynikov, 2004) have all been used. Each of these methods appears to guarantee that the estimated distribution will be a true probability distribution with all categories greater than or equal to zero and with individual probabilities that sum to one. In addition, all these methods appear to provide a theoretical basis for solutions that will be either maximum-likelihood estimates or at least convergent to a probability distribution.

However, all these methods are presented in a theoretical context that is useful in understanding the theory, but may not be suitable for actual application. Currently, most modelers will have an optimization program available. This note describes how, in a brief amount of time, the modeler can add a mixture model program to his collection of readily available programs—one that will estimate maximum-likelihood estimates for the mixture problem and will incorporate the experimenter's familiar optimization program.

To do this it is necessary to parameterize the estimated probabilities so that they are in fact guaranteed to constitute a probability distribution (i.e., $\hat{p}_j>0$, $\Sigma\hat{p}_j=1$). Two such parameterizations are the multinomial logit and a parameterization method described by Heifetz and Fujioka (1991). The trick with both parameterizations is that although the parameters that are actually estimated are unconstrained, these unconstrained estimates can be easily transformed to constrained probability estimates.

Materials and methods

Multinomial logit parameterization

Consider parameterizing the probabilities using the classic logit model:

$$p_j = \frac{r_j}{(1+\sum r_k)} \quad j,k = 1,\ldots,a-1 \text{ and}$$

$$p_a = \frac{1}{(1+\sum r_k)}.$$

The r_j can be guaranteed to be positive by defining $r_j=\exp(x_j)$. The parameters that are estimated are $x_j=\ln(r_j)$. In turn the $\{x_j\}$ are used to estimate the $\{r_j\}$, and then the $\{p_j\}$. Thus, $x_1, \ldots, x_{a-1}$ if are estimated on $(-\infty, \infty)$, the resulting $\{\hat{p}_j\}$ is guaranteed to be a probability distribution. The relationship between the $\{p_j\}$ and $\{r_j\}$ is a unique one-to-one mapping because given any $\{p_j\}$, $r_j=p_j/p_a$, for $j=1, \ldots, a-1$.

Heifetz and Fujioka parameterization

Heifetz and Fujioka (1991), for a tagged fish migration problem, used a somewhat similar parameterization that guaranteed estimated parameters would be probability distributions. Suppose a distribution consisted of categories with probabilities $\{p_j\}$. The Heifetz-Fujioka (H-F) parameterization would be

$$p_j = \frac{r_j}{\sum r_k}\left(1-\exp(-\sum r_k)\right)$$
$$j,k = 1,\ldots,a-1 \text{ and}$$

$$p_a = \exp(-\sum r_k)$$
$$\text{where } r_k > 0 \text{ for } k = 1,\ldots,a-1.$$

As above r_j, can be guaranteed to be positive by defining $r_j=\exp(x_j)$ and the parameters that are estimated are $x_j=\ln(r_j)$. Also as above, $\{x_j\}$ are used to estimate the $\{r_j\}$, and then the $\{p_j\}$. Thus, if $x_1, \ldots, x_{a-1}$ are estimated on $(-\infty, \infty)$, the resulting $\{\hat{p}_j\}$ is guaranteed to be a probability distribution. The relationship between the $\{p_j\}$ and $\{r_j\}$ is a unique one-to-one mapping because given any $\{p_j\}$, $r_j=-p_j\ln(p_a)/(1-p_a)$, for $j=1, \ldots, a-1$.

H-F probabilities as exploitation rates

Stock assessment modelers would recognize the formula

$$p_j = \frac{r_j}{\sum r_k}\left(1-\exp(-\sum r_k)\right)$$

as the exploitation rate formula where p_j takes the role of the exploitation rate (u_j) and r_j takes the role of the instantaneous mortality rate F_j. The formula $r_j=-p_j\ln(p_a)/(1-p_a)$ indicates that the instantaneous mortality rate can be solved in closed form, but this is generally not true because the natural mortality rate M is generally known as an instantaneous mortality rate rather than an exploitation rate.

Manuscript submitted 20 January 2005 to the Scientific Editor's Office.

Manuscript approved for publication 2 August 2005 by the Scientific Editor.

Fish. Bull. 104:303–305 (2006).

However, when exploitation rates are all known, as in some ecological modeling exercises (e.g., $u_j=C_j/N_j$, i.e., catch per population), then all instantaneous mortality rates can be solved in closed form.

Example: As an example, the logit and H-F parameterization can be used to estimate the distribution mixture problem for empirical length-at-age data. To describe the mixture problem for empirical length-at-age data, let

$i = 1, \ldots, n$ be the length category index;
$j = 1, \ldots, a$ be the age category index;

f_i = the observed length frequency, for which we wish to estimate the corresponding age distribution;
$\bar{f}_i = f_i/f$, where $f=\Sigma f_i$, the observed length distribution;
p_j = the unknown age distribution we wish to estimate from the observed length distribution;
q_{ij}=Prob$[i|j]$ = the observed distribution of length at age, estimated from age-length data;
$l_i = \Sigma_j p_j q_{ij}$ = length distribution estimated from an age distribution and observed length-at-age distributions.

Solutions to the empirical distribution mixture problem can be stated as solving for that age distribution $\{p_j\}$, which when combined with length i at age j, say $\{q_{ij}\}$, provides the "best" weighting to reproduce some length frequency $\{\bar{f}_i\}$. "Best" may be defined as

1 Maximum likelihood (Kimura and Chikuni, 1987):

$$L = const. + \sum_{i=1}^{n} f_i \log(l_i).$$

2 Minimum chi-square

$$\chi^2 = \sum_{i=1}^{n} (f_i - f l_i)^2 / (f l_i).$$

For both of these estimation problems, the estimated $\{\hat{p}_j\}$ distribution must be constrained to sum to one. The logit or H-F parameterization simplifies and unifies the estimation of $\{p_j\}$ for these or any other objective function. Any multivariate function optimization program, using these parameterizations, can generate estimated probability distributions whose components are positive and will be guaranteed to sum to one. Therefore any multivariate optimization program with these parameterizations can be used to estimate $\{p_j\}$.

Results

Greenland turbot (*Reinhardtius hippoglossoides*) length-at-age data were originally used to illustrate the iterated age-length key (IALK) (Kimura and Chikuni, 1987).

Table 1

Age distributions for Greenland turbot (*Reinhardtius hippoglossoides*) estimated by fitting length at age data to 1983 length-frequency data (Kimura and Chikuni, 1987). Except for the iterated age-length key (IALK) algorithm, all estimates were arrived at by using the logit or Heifetz-Fujioka (H-F) parameterizations which provided identical results.

Age (yr)	Maximum likelihood	IALK algorithm	Minimum χ^2
4	0.0353	0.0353	0.0377
5	0.1903	0.1903	0.1858
6	0.2281	0.2281	0.2229
7	0.1291	0.1291	0.1318
8	0.1125	0.1125	0.1117
9	0.0380	0.0380	0.0349
10	0.0525	0.0525	0.0554
11	0.0000	0.0000	0.0000
12	0.0332	0.0332	0.0383
13	0.0023	0.0023	0.0000
14	0.0204	0.0204	0.0235
15	0.0168	0.0168	0.0162
16	0.0289	0.0289	0.0248
17	0.0253	0.0253	0.0218
18	0.0680	0.0680	0.0719
19	0.0042	0.0042	0.0036
20	0.0152	0.0152	0.0198

We used the 1983 length-frequency data from that data set, along with the length-at-age data, to illustrate the methods of estimating mixture probabilities with the logit and H-F parameterizations (Table 1). Except for results from the IALK algorithm, all parameters were estimated by using an optimization program with the logit and H-F parameterizations (i.e., the latter two parameterizations gave identical results). Results also showed that maximum-likelihood estimates from either the logit or H-F parameterizations provided maximum-likelihood estimates identical to those estimated by using the IALK algorithm.

Discussion

It is probable that the logit and H-F parameterizations would provide nearly identical solutions for a given data set and objective function. If maximum likelihood solutions are unique, either parameterization should provide the maximum-likelihood estimates because the reparameterizations of probabilities are one-one mappings that lead to invariance properties when optimization is performed. The maximum-likelihood solutions will generally be unique when all $\hat{p}_j>0$ (Kimura and Chikuni, 1987). However, difficulties in searching in multivariate spaces, and limited computational precision may cause differences in the estimates.

Simple parameterizations, like direct estimates of $\{p_j\}$, will allow solutions that have negative probabilities. Constrained solutions, which allow maximum likelihood estimates and other types of estimates, will generally make these negative components of the probability distribution have zero values. Such solutions are boundary value solutions and may not be unique (Kimura and Chikuni, 1987). This is another reason it is difficult to claim unique solutions for the mixture problem.

From the modeling perspective, we illustrate the usefulness of reparameterization to impose mathematical constraints. In the context of the mixture problem the suggested parameterizations are reasonably transparent and allow the modeler to use familiar software. The reason we propose the methods described in this note is not that these methods provide superior estimates to those described in the literature, but that the procedure for estimation may actually be more straightforward and transparent for modelers more interested in solutions than in theory.

Because of its simplicity, effectiveness, and ready applicability to different objective functions, modelers may prefer optimization using the logit or H-F parameterizations to estimate probability distributions for the mixture problem. Another advantage of these reparameterizations is that they can be more generally applied, for example, to estimate geographic distribution in migration models (Heifetz and Fujioka, 1991; Shimada and Kimura, 1994).

Acknowledgments

We thank two anonymous referees for comments that helped us clarify our presentation.

Literature cited

Clark, W. G.
1981. Restricted least-squares estimates of age composition from length composition. Can. J. Fish. Aquat. Sci. 38:297–307.

Hasselblad, V.
1966. Estimation of parameters for a mixture of normal distributions. Technometrics 8:431–444.

Heifetz, J., and J. T. Fujioka.
1991. Movement dynamics of tagged sablefish in the northeast Pacific. Fish. Res. 11:355–374.

Hoenig, J. M., and D. M. Heisey.
1987. Use of a log-linear model with the EM algorithm to correct estimates of stock composition and to convert length to age. Trans. Am. Fish. Soc. 116:232–243.

Kimura, D. K., and S. Chikuni.
1987. Mixtures of empirical distributions: an iterative application of the age-length key. Biometrics 43:23–35.

Shimada, A. M., and D. K. Kimura
1994. Seasonal movements of Pacific cod, *Gadus macrocephalus*, in the eastern Bering Sea and adjacent waters based on tag-recapture data. Fish Bull. 92:800–816.

Troynikov, V. S.
2004. Use of weak quasi-solutions of the Fredholm first-kind equation in problems with scarce data. Appl. Math. Comput. 150:855–863.