CHAPTER 24

An Introduction to Statistical Algorithms Useful in Stock Composition Analysis

MICHAEL H. PRAGER AND KYLE W. SHERTZER

Population Dynamics Team, Center for Coastal Fisheries and Habitat Research, National Oceanic and Atmospheric Administration, Beaufort, North Carolina, USA

I. THE PROBLEM AND ITS TERMINOLOGY

In many fisheries, catches include fish that are conspecific but that originate in several spawning stocks. Because the population effects of fishing—and thus the choice of suitable management approaches—depend on which stock or stocks are harvested, estimates of stock composition of catches are needed. This need has given rise to the set of techniques often labeled *stock identification*. The focus of applications is usually on proportions in the harvest rather than on origin of individual fish; consequently, a more precise description of this work is *stock composition analysis*.

Stock Identification Methods

We define stock composition analysis as estimation of the stock composition of a mixed-stock sample (usually, some part of the harvest) taken from a known number J of component stocks. The proportion that originates in any given stock j is represented P_j. Thus $\Sigma_{j=1}^{J} P_j = 1$; within this constraint, any particular P_j may equal zero. The process by which the P_j are estimated constitutes the stock composition analysis.

Data used for such analyses are observations on characteristics of individual specimens; typical characteristics may include morphometrics, meristics, genetic characters, or chemical signatures (reviewed in Begg and Waldman, 1999). When we refer to characteristics here, we assume that they have been quantified in some reasonable way so that statistics (such as mean and variance) can be computed for the entire set (matrix) of observed characteristics. The probabilistic nature of the methods considered here is needed to account for overlap in the distribution of characteristics from different populations. When there is no overlap (as when using tags), the origin of each fish can be established with certainty, and much simpler methods can be used to define the stock composition of the catch.

This discussion also assumes the availability of a training sample of individuals whose stock membership is known. The training sample is used to fit a model by means of the investigator's choice of algorithm, a word used here to denote a statistical method or group of related methods. The fitted model is then used to estimate the stock composition of a mixed-stock sample (or samples) of the catch. The estimation can, but need not always, involve estimating the probability of stock membership of each individual in the mixed-stock sample. If in the course of estimation each individual is assigned membership in a particular stock, the method can be considered a classification method. Classification methods are a subset of all methods useful for stock composition analysis, because stock composition can be estimated without performing a classification.

The terminology of stock composition analysis is specific to fishery science, but the general problem is not. Analyzing characteristics of individual objects in a mixture to estimate the mixture's proportions is a general statistical problem known as finite mixture analysis. Constituent fish stocks in a mixed harvest are just one example of constituent classes or statistical populations of mixed objects; here, we tend to use the term *class* when describing algorithms generally, and *stock* when describing fisheries applications. Estimating stock composition is therefore a special case of the general problem of estimating mixing proportions.

We continue this chapter by introducing some algorithms useful for stock composition analysis. We then discuss issues involved in estimating performance of various algorithms, either in an absolute sense or relative to one another on a particular data set. We close with a few general comparative remarks.

II. ALGORITHMS

Many statistical algorithms have been proposed or used for stock composition analysis. We consider four—discriminant analysis, logistic regression, artificial neural networks, and finite mixture distributions—each of which may incorporate more than one observed characteristic. Variants of each algorithm exist, but since we focus on the conceptual basis of each algorithm, our treatment of such variants is generally brief.

A. Discriminant Analysis (DA)

Among classification schemes, discriminant analysis (McLachlan, 1992; Johnson and Wichern, 1998; Hastie et al., 2001) boasts the longest history. Its two most common forms are linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). Linear discriminant analysis was the first formal statistical method used for stock composition analysis (Hill, 1959), and the method has been used many times and in many variations. One of its advantages is the wide availability of well-tested and flexible software, as discriminant analysis forms an important component of most major statistics packages.

The fundamental assumption of LDA is that observed characteristics follow a multivariate normal distribution with common variance–covariance structure among stocks. If the characteristics vector is represented $\boldsymbol{X}$, we can write this distribution for stock j as $f_j(\boldsymbol{X}) \sim \text{MVN}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}_j$ is the mean characteristics vector for stock j and $\boldsymbol{\Sigma}$ is the common variance–covariance matrix. Typically, $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_j$, $j = \{1, \ldots, J\}$ are estimated from the training set, and those estimates are used in forming discriminant functions.

Classification in LDA is determined from stock-specific linear discriminant functions, computed from three types of information: an individual's characteristics vector $\boldsymbol{x}$; estimates of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}$; and a set of prior probabilities, or *priors*, p_j, $j = \{1, \ldots, J\}$. The priors are the analyst's estimates (which may be subjective) of the probabilities that a randomly chosen fish originates in each of the j stocks.

The discriminant function for stock j, evaluated for individual $\boldsymbol{x}$, is

$$\delta_j(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + \log(p_j) \tag{1}$$

where the notation $\boldsymbol{M}^T$ indicates the transpose of vector or matrix $\boldsymbol{M}$. The quantity $-\delta_j(\boldsymbol{x})$ measures the distance from individual $\boldsymbol{x}$ to the center of stock j (the function is conventionally in negative distance for computational reasons). In classification, each individual is assigned membership to the stock that maximizes $\delta_j(\boldsymbol{x})$, which is the stock with the closest center. This is also the stock with the largest posterior probability (eq. 3 of Pella and Masuda, this volume, Chapter 25).

The J discriminant functions thus define decision boundaries that classify an individual, from its observed characteristics, to the most likely stock of origin. The decision boundary between stocks j and k occurs where $\delta_j(\boldsymbol{x}) = \delta_k(\boldsymbol{x})$, and it is this decision boundary that, when solved for $\boldsymbol{x}$, is linear in the observed characteristics. If there are two measured characteristics, the decision boundary is a line; if three, a plane, and so on.

Quadratic discriminant analysis (QDA) is often considered preferable for problems in which the variance–covariance structure differs by class (Misra, 1985), as QDA does not assume equality of variance among classes (Kendall et al., 1983). However, estimates from QDA generally are of higher variance than those from LDA because of the additional parameters that must be estimated.

The quadratic discriminant function for stock j, evaluated for individual $\boldsymbol{x}$, is

$$\delta_j(\boldsymbol{x}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j) + \log(p_j) \qquad (2)$$

where $\boldsymbol{\Sigma}_j$ is the stock-specific variance–covariance matrix and $|\boldsymbol{\Sigma}_j|$ is its determinant. Classifications and posterior probabilities of class membership are computed as in LDA.

The decision boundary in QDA is defined as in LDA, but the resulting boundary is quadratic (curved) in the observed characteristics. For that reason, decision boundaries for LDA and QDA generally differ (Fig. 24-1).

In using either linear or quadratic discriminant analysis to estimate stock composition, one can proceed in two slightly different ways. The usual procedure, which we term *discrete classification*, is to classify each individual in the (mixed-stock) sample into the class for which its estimated membership probability is highest. The estimate of stock composition is then formed from the relative numbers of individuals classified into each class. In the second procedure, which we term *nondiscrete classification*, the probability of membership in a particular class is summed across all individuals. The estimate of stock composition is then obtained from the sums for each class divided by the total sample size.

As a simple example, consider a two-stock problem in which three fish are in the mixed-stock sample. Let the estimated probabilities of membership in class I for the three fish be {0.55, 0.45, 0.8}. The discrete estimate of mixing proportions would be 2/3 from class I and 1/3 from class II. The nondiscrete estimate would be 0.6 from class I and 0.4 from class II. The discrete estimate is obtained because two of the three fish are thought more likely to belong to class I; the nondiscrete is the mean of the three probabilities given.

Although discrete and nondiscrete classification usually produce similar estimates, it seems logical to prefer nondiscrete classification. There is no necessity to round estimated membership probabilities to whole numbers, as done in discrete classification, when the objective is to estimate mixing proportions.

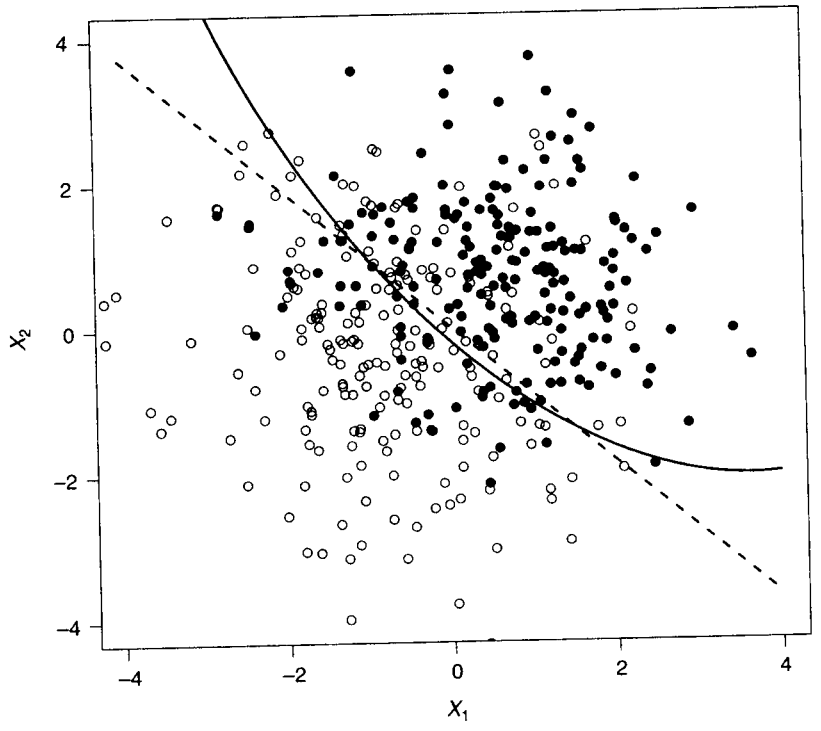

FIGURE 24-1. Example of decision boundaries between stock j (filled circles) and stock i (open circles) from linear (dotted line) and quadratic (solid line) discriminant analysis based on two observed characteristics (X_1, X_2). The prior probabilities are $p_i = p_j = 0.5$; the means are $\boldsymbol{\mu}_j = \begin{pmatrix} 0.6 \\ 0.75 \end{pmatrix}$ and $\boldsymbol{\mu}_i = \begin{pmatrix} -0.8 \\ -0.5 \end{pmatrix}$; the variance–covariance matrices are $\boldsymbol{\Sigma}_j = \begin{pmatrix} 1.5 & -0.25 \\ -0.25 & 1.25 \end{pmatrix}$ and $\boldsymbol{\Sigma}_i = \begin{pmatrix} 1.8 & -0.2 \\ -0.2 & 2.0 \end{pmatrix}$; and, with equal sample sizes, the pooled matrix $\boldsymbol{\Sigma}$ for the linear boundary is the mean of $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_j$.

Several other variants of discriminant analysis have been applied to stock identification; these include polynomial discriminant analysis (Cook and Lord, 1978), age-invariant discriminant analysis (Fabrizio, 1987), jackknife discriminant analysis (Small et al., 1998), and stepwise discriminant analysis (Palma and Andrade, 2002). Correction matrices, which can be computed from classification results on a test data set, are frequently used to improve mixture estimates from discriminant analyses (Cook and Lord, 1978; Pella and Robertson, 1978) and

might be used to correct estimates from other classification-based methods as well. Millar (1987) demonstrated that the use of classification with correction is a special case of maximum-likelihood finite mixture distribution methods (described below).

B. Logistic Regression

Logistic regression (Aldrich and Nelson, 1984; Hosmer and Lemeshow, 1989; Agresti, 2002) is a type of generalized linear model (McCullagh and Nelder, 1989). It was suggested for stock identification by Prager and Fabrizio (1990), who found the method promising. Its chief theoretical advantage is that it assumes neither multivariate normality of input data nor equality of variances and is appropriate for a wide variety of distributions (Kendall et al., 1983). It can also handle input data that are continuous, categorical, or a mix rather than continuous only, as in discriminant analysis. Logistic regression is applied most often to problems with a binary response, as when analyzing mixtures of two source stocks. But its use is not limited to binary problems, and indeed logistic regression has been applied to stock identification problems with more than two stocks (Prager and Fabrizio, 1990; Waldman et al., 1997).

In binary logistic regression, the response (Y_i) for fish i takes one of two values: $Y_i = 0$ implies membership in the first stock, and $Y_i = 1$ implies membership in the second. The probability that fish i with N measured characteristics $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$ belongs to the second stock is estimated by the continuous logistic function π,

$$P(Y_i = 1 \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i) = \frac{\exp(z_i)}{1 + \exp(z_i)} \tag{3}$$

where

$$z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_N x_{iN} \tag{4}$$

and the β's are parameters to be estimated (Fig. 24-2). The analysis defined by eq. 4 is called multiple logistic regression, which refers not to the number of stocks, but to analyzing more than one characteristic per fish (i.e., $N > 1$). A suitable transformation, accomplished by use of a link function, converts the model of eq. 3 into one that is linear. Several standard links exist for binary data, such as logit, probit, or log-log (Agresti, 2002); we present here the logit link because it is most often applied in polytomous logistic regression (more than two classes). The logit link is

$$\log\left(\frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_N x_{iN} \tag{5}$$

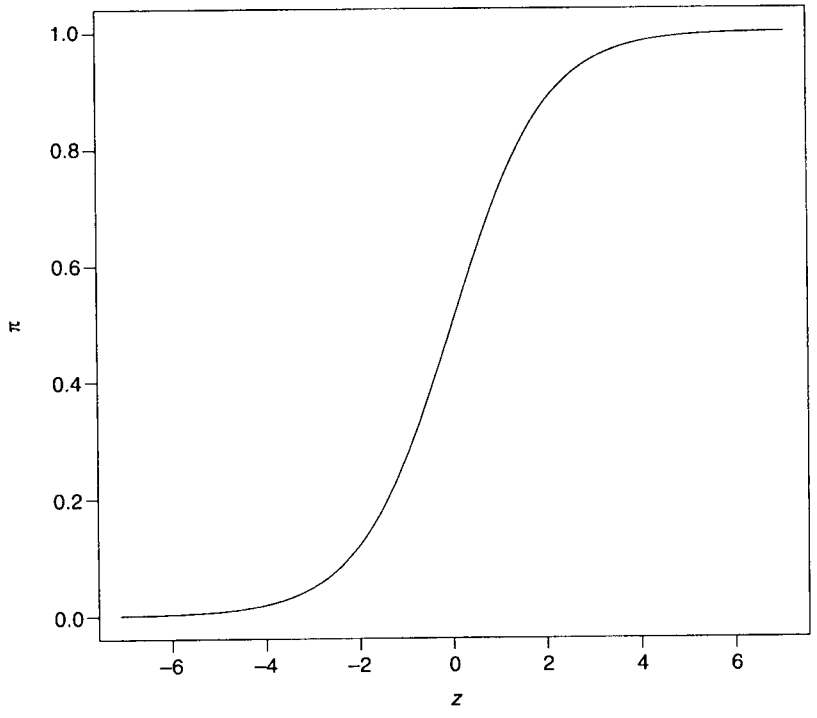

FIGURE 24-2. Logistic transformation shown in eq. 3. In logistic regression for stock identification, z is an unbounded linear combination of the measured characteristics. The value of the transformation π is the estimated probability of membership in the second of two stocks.

After transformation, errors are assumed to be distributed binomially, and parameters can readily be estimated by maximum likelihood using standard statistical software.

If there are more than two classes, binary logistic regression can be extended to polytomous logistic regression (Hosmer and Lemeshow, 1989; Agresti, 2002). The major difference is that now errors are assumed to be distributed multinomially rather than binomially. Classification among J stocks requires $J - 1$ link functions, which is no different from the binary case. For stock membership $j = 1, \ldots, J$,

$$P(Y_i = j \mid \boldsymbol{x}_i) = \pi_j(\boldsymbol{x}_i) = \frac{\exp(z_{ij})}{\sum_{h=1}^{J} \exp(z_{ih})} \tag{6}$$

where z_{ij} are analogous to eq. 4 but with parameters $\beta_{0j}, \beta_{1j}, \ldots, \beta_{Nj}$. The problem is constrained by the requirement that the probabilities of stock membership sum

to one. Given $J-1$ estimates of π_j, the Jth estimate must be $\pi_J(x_i) = 1 - \Sigma_{j=1}^{J-1}\pi_j(x_i)$. That constraint is implemented by defining parameters for the Jth stock to be zero (i.e., $\beta_{0J}, \beta_{1J}, \ldots, \beta_{NJ} = 0$). As a result, eq. 6 simplifies because $\exp(z_{iJ}) = 1$, and the π_j's can be estimated under transformation by a link function. As in the binary case, there are $J-1$ unique logit equations,

$$\log\left(\frac{\pi_j(x_i)}{\pi_J(x_i)}\right) = \beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + \cdots + \beta_{Nj}x_{iN} \qquad (7)$$

Computer programs for logistic regression are readily available in the major statistics software packages. In using such software, the analyst usually wishes to specify that the stock designations are nominal rather than ordinal values. Careful reading of the software's documentation may be needed to effect that choice, which is not always the default.

C. Artificial Neural Networks

The term *artificial neural network* (ANN) is not a precise one, but refers to a group of computational algorithms that sift through and combine many models to arrive at a model of optimum (in some sense) complexity (Ripley, 1996; Hastie et al., 2001). Unlike the other classification schemes presented here, ANNs are nonparametric. They require no assumptions about the distribution of data nor the particular functional relationship between model input and output. This can be a major advantage when such assumptions would be violated. Nonetheless, the success of such methods still depends on similarity of the data in the mixed-stock sample to the data in the training set.

Artificial neural networks have been developed in analogy to the structure of the human brain. Like the brain, an ANN consists of interconnected layers that process information provided by neurons. The input layer performs computations on the input data, and the results, along with a constant (bias), are then passed to a hidden layer. That procedure is iterated among a series of hidden layers until finally results are passed to the output layer. The number of hidden layers and the number of neurons in each can be adjusted to reflect the complexity of the problem. Once the architecture is established, values of network parameters are chosen as those that minimize some fitting criterion, a task usually accomplished with a learning algorithm.

Neural networks have proved useful in numerous fields, such as artificial intelligence, image compression, medical diagnosis, nondestructive testing, signal processing, and terrain classification, to name only a few. To our knowledge, the first published application in stock composition analysis was by Prager (1984, 1988) to estimate stock composition of striped bass and American shad. That study used

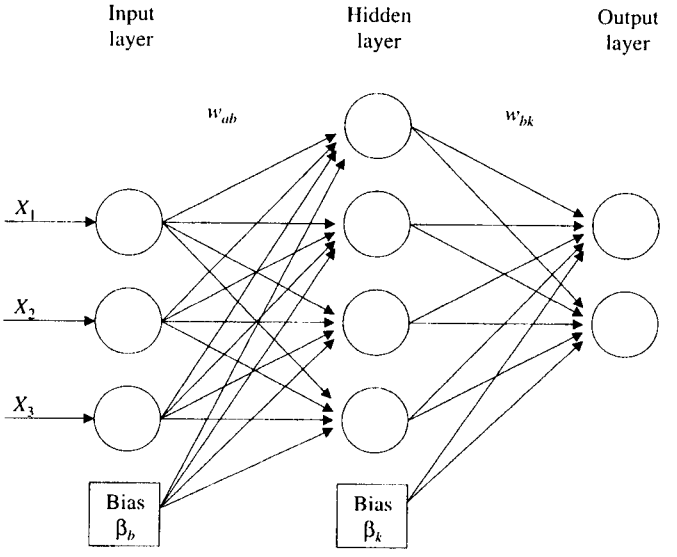

FIGURE 24-3. Example of an artificial neural network with a single hidden layer. The input variables (X_a) are the measure characteristics. Their weighted (w_{ab}) values plus a bias term (β_b) are processed by the hidden layer, and in turn, those results are processed similarly by the output layer. The final result is probabilities of stock membership, among two stocks in this example.

the group method of data handling (Ivakhnenko and Ivakhnenko, 1974), a type of neural network based on polynomials. More recent studies have compared the performances of ANNs and LDA and have found ANNs to be slightly more successful, at least on the particular data sets analyzed (Taylor and Beacham, 1994; Thorrold et al., 1998; Wells et al., 2000).

In the context of stock composition analysis, the input variables for an ANN are the measured characteristics; the output is the stock classification. Figure 24-3 illustrates a neural network with a single hidden layer that classifies a sample into two component stocks based on three characteristics. Each neuron in the hidden layer sums the weighted (w_{ab}) input signals (X_1, X_2, X_3) and adds a bias term (β_b). The result is then processed by a hidden-layer function (f_H) to produce an input signal for the output layer. The procedure is the same at the output layer, but with an output-layer function (f_O). Figure 24-3 can be expressed mathematically as

$$Y_k = f_O\left(\beta_k + \sum_b w_{bk} f_H\left(\beta_b + \sum_a w_{ab} X_a\right)\right) \tag{8}$$

where Y_k is the probability of membership in stock k. Any monotonic smooth function can be used for f_H or f_O, and they need not be the same, but the most popular by far is the logistic function of eq. 3.

Because of the complexity and specialized nature of the calculations, application of neural networks is generally made with specialized software. Because the methods are not standardized, different programs may offer different procedures and different results. For that reason, it may not be possible to duplicate existing results unless the same software is used, a situation that differs markedly from the other methods described here.

D. Finite Mixture Distribution Methods

Discriminant analysis, logistic regression, and neural networks can be considered classification-based algorithms because their focus is on an estimated classification of each individual (to stock), at least probabilistically. Those methods estimate, for each individual, the probability of membership in each class, and the estimates of composition—the desired results of fish stock composition analysis—are derived from the estimated membership probabilities of the individuals. The final set of algorithms discussed here, methods based on finite mixture distributions (FMD), does not share that focus on individuals. (Here *finite* refers to the number of classes in the mixture.) Although the probability of an individual's group membership can be estimated from finite mixture methods, the primary focus is on estimating the composition of a mixed sample (mixture distribution). This difference in focus is important.

Maximum-likelihood estimation of finite mixture distributions has been the subject of several books in the statistical literature (Wolfe, 1970; Everitt and Hand, 1981). The methods were introduced to the fisheries literature to separate size compositions into age classes (Cassie, 1954; Bhattacharya, 1967). Application to stock composition analysis came later (Fournier et al., 1984; Pella and Milner, 1986; Millar, 1987; Wood et al., 1987). The methods are simplest to apply if one assumes that characteristics follow a multivariate normal distribution with a common covariance matrix among classes, the same assumption used in linear discriminant analysis. However, FMD methods can be adapted to a wide variety of distributions and can accommodate unequal covariance matrices (e.g., DeVries et al., 2002).

A finite mixture distribution (f) describes the distribution of a vector ($\boldsymbol{x}$) of N observed characteristics. The mixture distribution f can be expressed as a weighted sum of its J component probability distributions g_j, $j = 1, \ldots, J$ (where as before J is the number of stocks). For example, a mixture of multivariate normal distributions with unequal variances can be written,

$$f(\boldsymbol{x} \mid \boldsymbol{p}, \boldsymbol{\mu}, \Sigma) = \sum_{j=1}^{J} p_j g_j(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \Sigma_j) \tag{9}$$

Here $\boldsymbol{p} = (p_1, p_2, \ldots, p_{J-1})$ are the $J - 1$ independent mixing proportions, or weights, of the component distributions such that $0 < p_j < 1$ and $p_J = 1 - \Sigma_{j=1}^{J-1} p_j$. The $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the mean vector and variance–covariance matrix of the jth component distribution.

In a stock composition analysis, the mixed-stock sample is from the mixture distribution (f) of characteristics. The goal is to estimate mixing proportions (p_j); to do so, one must first estimate (from the training set) the parameters of the component distribution (g_j) of each constituent stock. The mixture distribution observed in the mixed-stock sample depends both on the parameters (shapes) of the stock-specific component distributions and on the mixing proportions in the mixed-stock sample (Fig. 24-4).

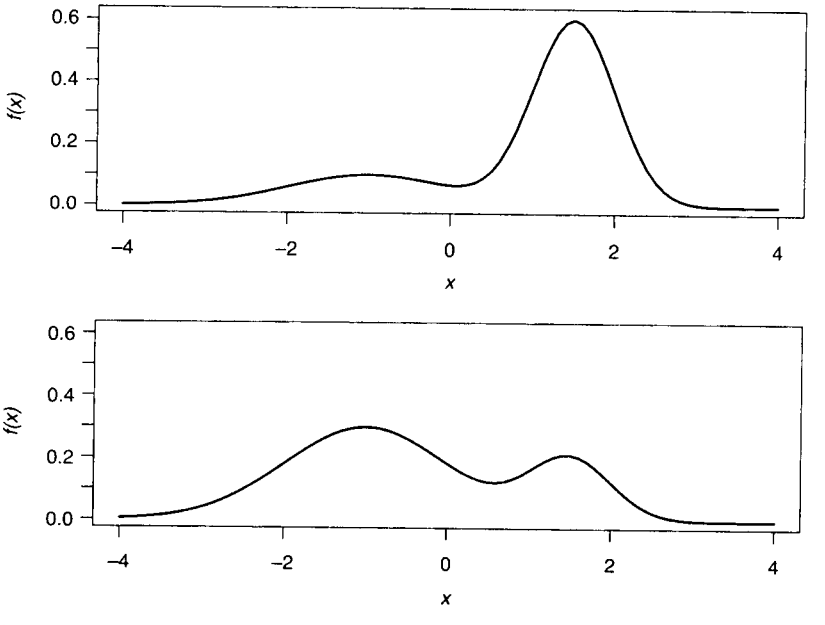

FIGURE 24-4. Mixture distributions with the same component distributions in different proportions. Each mixture is of two univariate normal distributions: stock one with $\mu_1 = -1.0$, $\sigma_1 = 1.0$, and stock two with $\mu_2 = 1.5$, $\sigma_2 = 0.5$. The mixing proportions are $p_1 = 0.25$, $p_2 = 0.75$ in the top panel, and $p_1 = 0.75$, $p_2 = 0.25$ in the bottom panel.

Practical Application of FMD Methods

Software for mixture problems seems less readily available than software for fitting discriminant functions, logistic regressions, and even neural networks, a situation that may have slowed the wide adoption of FMD methods in fisheries. However, R. Millar has developed Fortran software, HISEA, that, as of the date of writing, was available on the World Wide Web. (http://www.stat.auckland.ac.nz/~millar/mixedstock/code.html). Millar's software implements the FMD method under the assumption of multivariate normality with constant variance. Other suitable software (e.g., EMMIX by G. McLachlan) may also be freely available.

It may not be obvious that commercial software for discriminant analysis can also be used to obtain maximum-likelihood FMD estimates of stock composition. Thus, any analyst with access to standard statistical packages can explore the properties of FMD estimates.

To proceed in this way, it is simplest to assume that the measured characteristics are multivariate normal with a common covariance matrix across stocks. Under that assumption, the equations for LDA and FMD are identical, with the mixing proportions to be estimated by FMD corresponding to the the correct (but unknown) priors in LDA. By using an EM algorithm (Dempster et al., 1977), maximum-likelihood estimates of the mixing proportions can be obtained under the FMD model. As noted by Millar (1987), ". . . in constructing a classification rule, one is actually doing all of the work required to construct the likelihood function, so from there it would be a matter of simply running a maximization program to obtain the maximum likelihood estimates." The procedure is as follows:

1. Fit a linear discriminant function to the training sample.
2. Obtain a starting guess for the priors. In the absence of other information, one can use equal priors, that is, set $p_j = 1/J$ for all j.
3. Using the current priors and the discriminant function fit in step 1, make a nondiscrete estimate of the mixture proportions of the mixed-stock sample.
4. Revise the priors to equal the current estimated mixture proportions.
5. Repeat steps 3 and 4 until the composition estimates converge.

Without doubt, this procedure is more tedious than using software written specifically for the estimation of mixing proportions. However, if such software is not readily available, or if an investigator wishes to take the first steps into using FMD methods, this iterative procedure may prove useful.

III. THE IMPORTANCE OF PRIOR KNOWLEDGE

Each classification-based method described above (discriminant analysis, logistic regression, neural networks) either makes implicit use of properties of the training sample or requires additional assumptions to estimate stock composition. In the case of discriminant analysis, the composition estimates are conditional on the prior probabilities of stock membership (the p_i in eq. 1) specified by the analyst. Perhaps because standard software offers defaults for these priors, their importance is often overlooked. Nonetheless, the reliance on priors—or equivalent information from the training sample—is a major limitation of the classification-based methods. For that reason, we emphasize here the role of priors in discriminant analysis and the role of analogous information in other classification algorithms.

A. Priors and Discriminant Analysis

Priors, as used by discriminant analysis, are a priori estimates of the probability that an individual in the mixed-stock sample is a member of each component stock. By *a priori*, we mean that the individual is chosen at random from the mixed sample and that nothing further is known about it: its characteristics have not yet been observed. If we assume that the stocks are present in the mixed-stock sample in proportion to their presence in the mixture under consideration (except for sampling error), the paradox involved in using discriminant analysis for this problem becomes clear. The priors, which are required to make an estimate, are precisely what we are trying to estimate: the relative contribution of stocks in the mixture.

This paradox does not occur in some other fields that use classification methods because the structure of their questions is fundamentally different. For example, a typical medical application might be to estimate the probability of a patient's contracting some disease, given observations about general health and family history. In that example, the proportion of the general population that will contract the disease is well known and can be used as the prior. The focus of such a study is estimation about the individual. In such applications, reliable priors are readily available, and classification methods are appropriate. Notably, their use does not entail the same paradox as in stock composition studies.

In discriminant analyses, priors are specified explicitly, and composition estimates cannot be made without them. Statistics packages commonly offer two simple tactics for generating priors. The first tactic is to base priors on the proportions observed in the training set. Under that approach, the prior for a given stock is set equal to its relative predominance in the training sample. The unstated

assumption is that the composition of the training sample is a good estimate of the composition of the mixed population. The second tactic is to set the priors equal: with three stocks, the prior for each would be set to 1/3. Neither tactic has a valid theoretical basis, and neither is particularly likely to result in accurate priors. When priors are inaccurate, resulting estimates of mixture proportions are biased toward the priors, although correction matrices may reduce such biases.

Frequently, studies that use discriminant analysis provide estimates of classification error rates. In interpreting such estimates, it is important to understand that the error rates of an uncorrected discriminant estimator depend on the actual (and unknown) mixed-stock composition. Such error rates are generally smallest when the priors are accurate (i.e., when they correspond to the actual stock composition) and can be much higher when the priors are inaccurate. Reported error rates can also be biased low if the training-set data have been used to estimate the error rates.

B. Priors and Logistic Regression

In logistic regression, explicit priors are not specified by the analyst. However, a logistic regression estimator is derived from the distributions observed in the training sample. It therefore provides the least biased estimates when the mixed-class sample is of the same composition as the training sample. To translate the preceding into fisheries language, a logistic-regression estimator will be biased toward the stock composition of its training sample and will be increasingly inaccurate as the composition of the mixed-stock sample differs from that of the training sample. The use of correction matrices to reduce bias in mixing estimates from logistic regression seems feasible. However, we do not know that the specific subject has been studied.

Unfortunately, the bias just described was overlooked by Prager and Fabrizio (1990). When an adjustment was made to their experimental procedure to compensate, the same authors found the performance of logistic regression to be similar to that of linear discriminant analysis, even on nonnormal data (M. H. Prager, C. D. Jones, and M. C. Fabrizio, unpublished study).

C. Priors and Neural Networks

Estimation of stock composition by neural networks is also conditional on the composition of the training sample. Therefore, bias of these estimators should be similar to that from logistic regression. The ultimate performance of such a method also depends on the particular data and specific method involved. As

with logistic regression, it seems that the use of correction matrices with neural networks has not yet been studied, at least in fishery science.

D. Priors and FMD Methods

Finite mixture distribution methods do not require priors, implicit or explicit. It stands to reason that, in most cases, they should produce estimates with lower bias than uncorrected classification-based algorithms. This may be accompanied by increased variance.

IV. DISCUSSION

The four methods presented here are familiar techniques in stock composition analysis. However, other classification methods exist and may hold added promise for fisheries. One nonparametric technique is tree-based regression (often called CART), applied to stock classification by Weigel et al. (2002). A quite different approach is analysis of tagging data for stock composition (Schwartz, this volume, Chapter 28).

Evaluation of statistical algorithms for stock composition analysis is not a simple task, whether undertaken on simulated or real data. What has been underemphasized in some such evaluations is how an estimator performs as the composition of the mixed-class sample varies from that of the training sample. One would expect the error of uncorrected discriminant analyses to become considerably worse as the true composition varies from the priors. A similar deterioration in performance of uncorrected logistic regression or neural network estimates will occur as the true composition varies from that of the training sample. The error rate of FMD methods may deteriorate as the composition of the mixed-stock sample becomes quite different from that of the training sample, but variation in the composition of the mixed-stock sample does not constitute violation of a major assumption, as it does with the classification-based methods.

The analyst should always be aware of the chosen algorithm's assumptions and the degree to which they are met. Gray (1994) showed that in FMD applications, model misspecification can lead to poor estimation. For example, if unequal variances are assumed equal, or if skewed distributions are assumed normal, FMD results may be strongly biased (Gray, 1994). Similar problems can arise with any algorithm if assumptions are violated. Moreover, bias and variance are statistical properties, and there is no assurance that a specific algorithm will be more accurate or precise than another in a particular application. To some degree, properties of algorithms in specific applications can be examined through Monte Carlo simulation studies.

From the preceding, it can be seen that for the stock composition problem in fisheries, uncorrected methods appear least desirable, and FMD methods appear more appropriate than methods based on classification, provided assumptions are met. Millar (1990) concluded that corrected classification estimators are as useful as FMD methods when the number of stocks is small (two or three), but recommended use of direct maximum-likelihood estimation (FMD methods) for more complex problems. We concur with that recommendation, and we further recommend that when error rates are estimated, they should be reported based on a range of priors (or their equivalent), if used, and over a range of (simulated) stock compositions.

ACKNOWLEDGMENTS

Correspondence and conversations with J. Pella helped to clarify the relationship between discriminant analysis and the finite mixture problem. Insight was also gained during an unpublished simulation study conducted with M. Fabrizio and C. D. Jones. We thank R. Millar for sharing his HISEA software, and we thank D. Ahrenholz, P. Hanson, and J. Waters for reviewing the manuscript. This work was supported by the Virginia Sea Grant Program and the Southeast Fisheries Science Center of the U.S. National Marine Fisheries Service through the NOAA Center for Coastal Habitat and Fisheries Research.

REFERENCES

Agresti, A. 2002. Categorical Data Analysis, 2nd Ed. Wiley, New York. 734 pp.

Aldrich, J. H. and Nelson, F. D. 1984. Linear probability, logit, and probit models. Sage University Paper series on Quantitative Applications in the Social Sciences, 07–045. Sage Publications, Beverly Hills, CA.

Begg, G. A. and Waldman, J. R. 1999. An holistic approach to fish stock identification. Fish. Res. 43: 35–44.

Bhattacharya, C. G. 1967. A simple method of resolution of a distribution into Gaussian components. Biometrics 23: 115–135.

Cassie, R. M. 1954. Some uses of probability paper in the analysis of size frequency distributions. Austral. J. Mar. Freshw. Res. 5: 513–522.

Cook, R. C. and Lord, G. E. 1978. Identification of stocks of Bristol Bay sockeye salmon, *Oncorynchus nerka*, by evaluating scale patterns with a polynomial discriminant method. Fish. Bull. 76: 415–423.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc., Series B 39: 1–38.

DeVries, D. A., Grimes, C. B., and Prager, M. H. 2002. Using otolith shape analysis to distinguish eastern Gulf of Mexico and Atlantic Ocean stocks of king mackeral. Fish. Res. 57: 51–62.

Everitt, B. S. and Hand, D. J. 1981. Finite Mixture Distributions. Chapman & Hall, London. 143 pp.

Fabrizio, M. C. 1987. Growth-invariant discrimination and classification of striped bass stocks by morphometric and electrophoretic methods. Trans. Am. Fish. Soc. 116: 728–736.

Fournier, D. A., Beacham, T. D., Riddell, B. E., and Busack, C. A. 1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. Can. J. Fish. Aquat. Sci. 41: 400–408.

Gray, G. 1994. Bias in misspecified mixtures. Biometrics 50: 457–470.

Hastie, T., Tibshirani, R., and Friedman, J. 2001. The Elements of Statistical Learning. Springer, New York. 552 pp.

Hill, D. R. 1959. Some uses of statistical analysis in classifying races of American shad (*Alosa sapidissima*). Fish. Bull. 59: 269–286.

Hosmer, D. H. and Lemeshow, S. 1989. Applied Logistic Regression. Wiley, New York. 307 pp.

Ivakhnenko, A. G. and Ivakhnenko, N. A. 1974. Long-term prediction of random processes by GMDH algorithms using the unbiasedness criterion and balance-of-variables criterion. Sov. Autom. Control 7(4): 40–45.

Johnson, R. A. and Wichern, D. W. 1998. Applied Multivariate Statistical Analysis, 4th Ed. Prentice Hall, Upper Saddle River. 816 pp.

Kendall, M., Stuart, A., and Ord, J. K. 1983. The Advanced Theory of Statistics, Vol. 3, 4th Ed. Charles Griffin, London. 780 pp.

McCullagh, P. and Nelder, J. A. 1989. Generalized Linear Models, 2nd Ed. Chapman & Hall, London. 532 pp.

McLachlan, G. J. 1992. Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York. 544 pp.

Millar, R. B. 1987. Maximum likelihood estimation of mixed stock fishery composition. Can. J. Fish. Aquat. Sci. 44: 583–590.

Millar, R. B. 1990. Comparison of methods for estimating mixed stock fishery composition. Can. J. Fish. Aquat. Sci. 47: 2235–2241.

Misra, R. K. 1985. Quadratic discriminant analysis with covariance for stock delineation and population differentiation: a study of beaked redfishes (*Sebastes mentella* and *S. fasciatus*). Can. J. Fish. Aquat. Sci. 42: 1672–1676.

Palma, J. and Andrade, J. P. 2002. Morphological study of *Diplodus sargus*, *Diplodus puntazzo*, and *Lithognathus mormyrus* (Sparidae) in the Eastern Atlantic and Mediterranean Sea. Fish. Res. 57: 1–8.

Pella, J. J. and Milner, G. B. 1986. Use of genetic marks in stock composition analysis. In N. Ryman and F. Utter (eds.), Population Genetics and Fishery Management. University of Washington Press, Seattle, pp. 247–276.

Pella, J. J. and Robertson, T. L. 1978. Assessment of composition of stock mixtures. Fish. Bull. 77: 387–398.

Prager, M. H. 1984. The Group Method of Data Handling: Applications in Oceanography and Fishery Science. Ph.D. dissertation, University of Rhode Island, Kingston.

Prager, M. H. 1988. Group method of data handling: a new method for stock identification. Trans. Am. Fish. Soc. 117: 290–296.

Prager, M. H. and Fabrizio, M. C. 1990. Comparison of logistic regression and discriminant analyses for stock identification of anadromous fish, with application to striped bass (*Morone saxatilis*) and American shad (*Alosa sapidissima*). Can. J. Fish. Aquat. Sci. 47: 1570–1577.

Ripley, B. D. 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge. 403 pp.

Small, M. P., Withler, R. E., and Beacham, T. D. 1998. Population structure and stock identification of British Columbia coho salmon, *Oncorhynchus kisutch*, based on microsatellite DNA variation. Fish. Bull. 96: 843–858.

Taylor, E. B. and Beacham, T. D. 1994. Population structure and identification of North Pacific Ocean chum salmon (*Oncorhynchus keta*) revealed by an analysis of minisatellite DNA variation. Can. J. Fish. Aquat. Sci. 51: 1430–1442.

Thorrold, S. R., Jones, C. M., Swart, P. K., and Targett, T. E. 1998. Accurate classification of juvenile weakfish *Cynoscion regalis* to estuarine nursery areas based on chemical signatures in otoliths. Mar. Ecol. Prog. Ser. 173: 253–265.

Waldman, J. R., Richards, R. A., Schill, W. B., Wirgin, I., and Fabrizio, M. C. 1997. An empirical comparison of stock identification techniques applied to striped bass. Trans. Am. Fish. Soc. 126: 369–385.

Weigel, D. E., Peterson, J. T., and Spruell, P. 2002. A model using phenotypic characteristics to detect introgressive hybridization in wild westslope cutthroat trout and rainbow trout. Trans. Am. Fish. Soc. 131: 389–403.

Wells, B. K., Thorrold, S. R., and Jones, C. M. 2000. Geographic variation in trace element composition of juvenile weakfish scales. Trans. Am. Fish. Soc. 129: 889–900.

Wolfe, J. H. 1970. Pattern clustering by multivariate mixture analysis. Multivar. Behav. Res. 5: 389–398.

Wood, C. C., McKinnell, S., Mulligan, T. J., and Fournier, D. A. 1987. Stock identification with the maximum-likelihood mixture model: sensitivity analysis and application to complex problems. Can. J. Fish. Aquat. Sci. 44: 866–881.

Classic
Analys
Individ
Popula
of Mixt

JEROME PELLA A
United States Departm
National Marine Fishe

Stock Identification Methods