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.

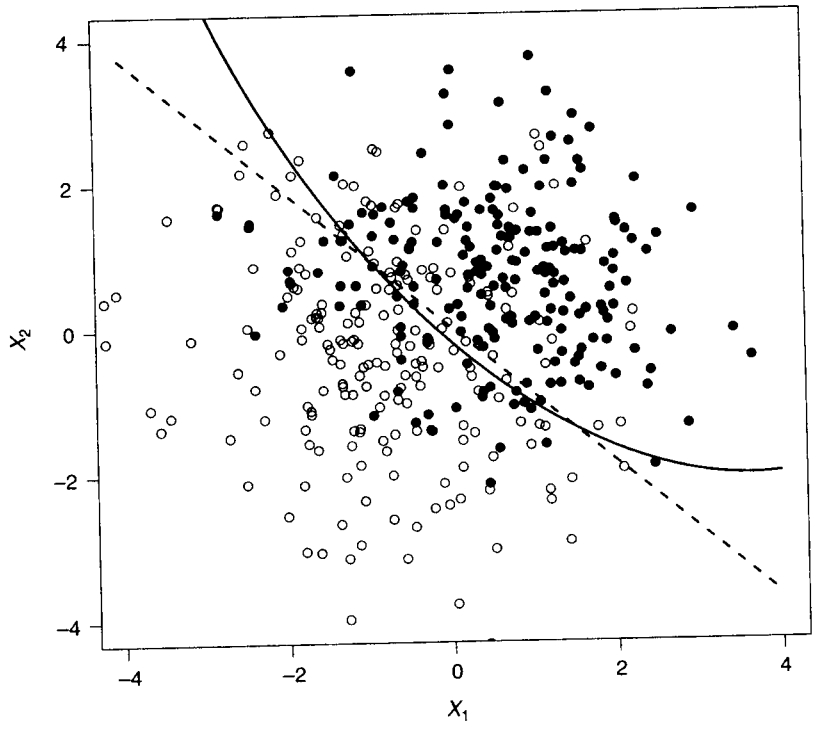

FIGURE 24-1. Example of decision boundaries between stock j (filled circles) and stock i (open circles) from linear (dotted line) and quadratic (solid line) discriminant analysis based on two observed characteristics (X_1, X_2). The prior probabilities are $p_i = p_j = 0.5$; the means are $\boldsymbol{\mu}_j = \begin{pmatrix} 0.6 \\ 0.75 \end{pmatrix}$ 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

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

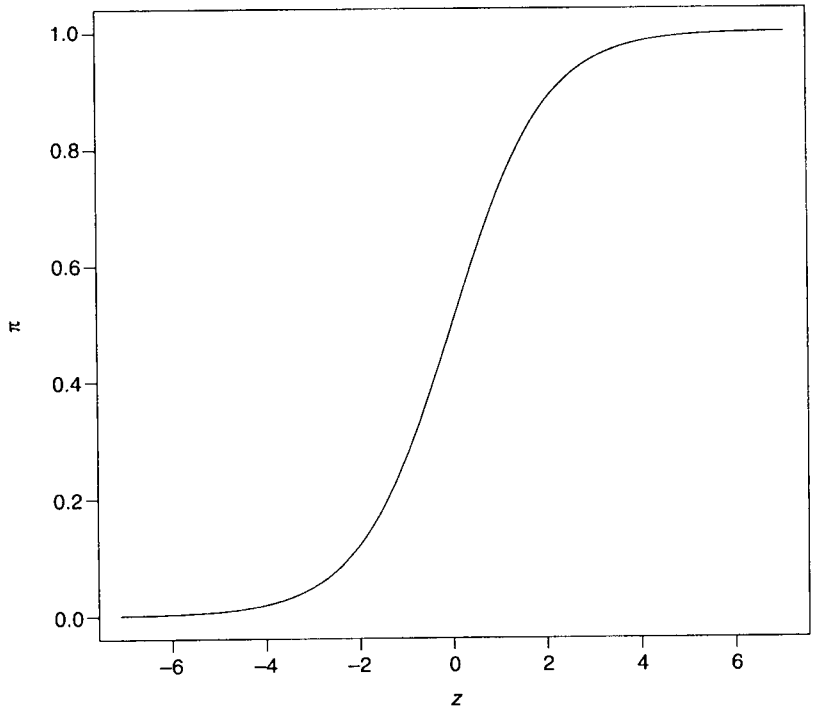

FIGURE 24-2. Logistic transformation shown in eq. 3. In logistic regression for stock identification, z is an unbounded linear combination of the measured characteristics. The value of the transformation π is the estimated probability of membership in the second of two stocks.

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

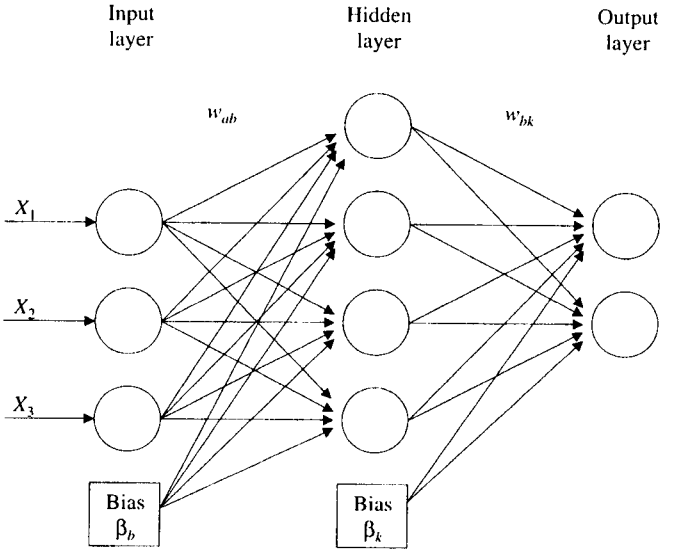

FIGURE 24-3. Example of an artificial neural network with a single hidden layer. The input variables (X_a) are the measure characteristics. Their weighted (w_{ab}) values plus a bias term (β_b) are processed by the hidden layer, and in turn, those results are processed similarly by the output layer.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

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