Vol. 329: 301–305, 2007 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published January 11

NOTE

Discriminant function analysis in marine ecology: some oversights and their solutions

J. Wilson White*, Benjamin I. Ruttenberg

Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, California 93106, USA

ABSTRACT: Marine ecologists commonly use discriminant function analysis (DFA) to evaluate the similarity of distinct populations and to classify individuals of unknown origin to known populations. However, investigators using DFA must account for (1) the possibility of correct classification due to chance alone, and (2) the influence of prior probabilities of group membership on classification results. A search of the recent otolith chemistry literature showed that these two concerns are sometimes ignored, so we used simulated data sets to explore the potential pitfalls of such oversights. We found that when estimating reclassification success for a training data set, small sample sizes or unbalanced sampling designs can produce remarkably high reclassification success rates by chance alone, especially when prior probabilities are estimated from sample size. When using a training data set to classify unknown individuals, maximum likelihood estimation of mixture proportions and group membership afforded up to 20% improvement over DFA with uninformative priors when groups contributed to the sample unequally. Given these results, we recommend the use of (1) randomization tests to estimate the probability that reclassification success is better than random, and (2) maximum likelihood estimation of mixture proportions in place of uninformative priors.

KEY WORDS: Classification · Discriminant function analysis · Jackknife reclassification success · Null hypothesis · Prior probabilities · Randomization test · Otolith chemistry

INTRODUCTION

In recent years, marine ecologists have relied increasingly on multivariate statistics to elucidate structure in their data sets. For example, somatic morphometrics are frequently used to discriminate fish stocks (Cadrin 2000) and a growing subfield of marine ecology uses chemical signatures deposited in calcified structures to infer patterns of movement between water masses (Campana 2005). Discriminant function analysis (DFA) is a popular statistical tool in these studies, partly because it can classify individuals of unknown origin into groups using a discriminant function (DF) generated from a training data set composed of individuals of known origin (see McGarigal et al. 2000 for an accessible mathematical treatment). However, DFA has limitations that are sometimes overlooked by practitioners. First, DFA will classify some samples correctly by chance alone, so the performance of a DFA must be evaluated against that chance success rate. Second, the assignments generated by DFA can be strongly affected by the prior probabilities of group membership, so poor estimation of these priors may undermine classification accuracy.

The classification accuracy of DFA is commonly evaluated by leave-one-out cross-validation, also called jackknife reclassification. This procedure omits 1 individual from the data set, recalculates the DF, and assigns the omitted individual to a group using the new DF (Lachenbruch & Mickey 1968). This process is then repeated for every sample in the data set. Since actual group membership is known, the fraction of samples correctly reassigned to their respective groups can be calculated; this is the jackknife reclassification success rate.

*Email: w_white@lifesci.ucsb.edu

A low jackknife value suggests that the DF is unable to classify samples accurately. However, some success is expected by chance even if there are no real differences among groups, the amount of which depends on the number of groups being compared, *g*: if groups are evenly represented in the data set, the null expectation for reclassification success should be 1/*g*. However, an unbalanced data set will deviate from this expectation; if 1 group dominates the data set, the DF may reclassify many samples correctly by chance even if no real differences exist among groups. This issue necessitates some sort of chance-corrected estimate of the null expectation for reclassification success (McGarigal et al. 2000). Unfortunately, the null expectations for complex data sets are often non-intuitive. Even when the null expectation is obvious, a direct comparison to the jackknife value is hindered by the high variance of the jackknife estimator at low sample sizes (Glick 1978). Clearly investigators need a reliable method for testing the null hypothesis that reclassification success is better than that expected by chance.

Wastell (1987) and Solow (1990) advocated the use of a randomization test for this purpose. While some fisheries scientists have applied this technique (Cadrin 2000), its utility has gone unheeded by many marine ecologists and in particular by otolith chemistry investigators. To highlight this problem, we performed a search of the ISI (Institute for Scientific Information) Web of Science database using the keywords 'otolith' and 'chemistry.' The search returned 81 articles between the years 2000 and 2005, of which 65 were empirical and 30 used DFA. None of these used a randomization technique to test the no-better-than-random null hypothesis, and only one (Wells et al. 2000) explicitly considered the 1/*g* reclassification success expectation. Several authors used other means to assess the reliability of their DFA results, such as complementary assignment techniques (e.g. artificial neural networks; Thorrold et al. 1998) or Cohen's kappa statistic (DeVries et al. 2002), although the latter is only appropriate for use with hold-out reclassification (see McGarigal et al. 2000 for details). Nonetheless, while most ecologists rarely report summary statistics (e.g. differences among sample means) without some estimate of the probability of obtaining those values by chance (e.g. p-values), authors regularly report reclassification success results without such supporting statistics.

It is important to clarify that the null hypothesis being tested here (no-better-than-random reclassification) is not equivalent to the parametric null hypothesis of no difference between the true group population means. The latter hypothesis is properly tested with multivariate analysis of variance (MANOVA), not DFA. Some otolith studies use MANOVA to ensure groups are different before proceeding with DFA (e.g. Thorrold et al. 1998); this worthwhile practice provides assurance that classification is possible but does not directly evaluate reclassification success.

A second issue confronting DFA users is the estimation of prior probabilities. These *a priori* estimates of the probability of membership in each group are incorporated into the DFs used for classification, and an arbitrary assignment of priors can result in similarly arbitrary classifications (Williams 1983). When reporting jackknife reclassification success, authors generally use non-informative uniform priors or let priors be proportional to group sample sizes. Depending on the evenness of sampling effort, this choice could greatly affect the reclassification success. When classifying truly unknown individuals, the choice of priors becomes crucial because the DF is dependent on the probability of an unknown belonging to a given group. At the same time, estimating priors is difficult because we rarely know the relative contribution of different groups to the pool of unknowns. For these reasons, the choice of priors is a contentious topic. Ideally, independent ancillary data such as relative stock abundances or flow regimes affecting fish movement would be fashioned into prior probabilities. In the absence of such information, maximum likelihood (ML) methods can be used in place of DFA to simultaneously estimate mixture proportions (i.e. posterior probabilities of group membership) and group assignments (Millar 1987). Because DFA can be cast as a likelihood-ratio method when the data are multivariate normal (Williams 1983), ML classifications are equivalent to those produced by DFA if the *a posteriori* ML mixture proportions were used as priors. Maximum likelihood methods are commonly used in the otolith literature to supplement inferences from DFA (Campana et al. 1999, 2000, Thorrold et al. 2001, Gillanders 2002, Wells et al. 2003). However, some investigators prefer to classify unknowns using DFA with uniform priors, and the difference in classification success between this method and ML classification is unknown.

Here we use a combination of empirical and simulated data to emphasize the potential problems of ignoring chance classifications and mis-estimating prior probabilities. We describe the proper method for testing the null hypothesis of no-better-than-random reclassification success and evaluate the success of the ML solution for the prior probability dilemma. We also provide Matlab code for both procedures.

METHODS AND RESULTS

Chance classification success. To determine the probability of obtaining a given jackknife value due to chance alone, one could generate the distribution of expected jackknife values if no differences exist among

groups. Comparing the observed value to this distribution will allow the investigator to estimate the probability that the observed jackknife value was drawn from this distribution, i.e. the probability (p-value) that this result was obtained by chance alone. The distribution of null jackknife values can be approximated by randomization: each individual is assigned to a group at random, then the jackknife reclassification success is calculated for the new, randomized data set. Repeating this process many times (the simulations presented here used 1000 randomizations) approximates the distribution of null jackknife values.

We used this method to generate some general guidelines for the interpretation of jackknife reclassification success rates in a range of scenarios. First we evaluated the effects of number of groups and overall sample size on DFA success by calculating the null jackknife reclassification success value for linear DFA using a simulated data set with 3 predictor variables (using more variables does not affect the results) and 2 to 4 groups with equal sample sizes ranging from 10 to 130 per group. Predictor variable values were drawn from the standard multivariate normal distribution ($\boldsymbol{\mu}_i = [0\,0\,0]$ for each group i, $\Sigma = \boldsymbol{I}$; our results hold with other covariance matrices) to create a data set with no differences among groups. The results of these simulations confirmed that null expectations of jackknife reclassification success are well approximated by $1/g$ for balanced samples (Fig. 1A). The variance associated with the null expectation was high for small sample size ($n < \sim 30$ per group) but decreased appreciably as sample size increased.

Choice of prior probabilities for jackknife reclassification estimator. Because real data sets often vary in sample size among groups, we also explored the change in null jackknife reclassification success as sample sizes become more unequal. For this simulation, we used two groups, varied the ratio of group sample size from 1:1 to 4:1, held total sample size constant at either 100 or 200, and generated multivariate normal data sets in the same way as above. In such cases there are 2 ways to calculate the reclassification success rate: with uniform prior probabilities or using 'empirical' priors equal to the relative sample sizes of the groups. When sample sizes are unbalanced, the latter method produces higher null reclassification success values than the former.

With either set of priors there was a monotonic increase in null jackknife values above the $1/g$ expectation with increasing inequality among group sample sizes (Fig. 1B). This increase was modest when uniform priors are used (an increase from 50% with equal sample size to 56% with a 4:1 ratio of sample sizes and n = 100) but extreme with empirical priors. When n = 200 with a 4:1 ratio of sample sizes, reclassification success must be >80% to exceed null expectations. This deviation from $1/g$ appears to result from increased chance reclassification success for the larger group. With both sets of priors, increasing sample size reduces the variance in the null expectation but has minimal effect on the deviation from $1/g$.

Classification of unknowns without informative priors. Here we used data from an investigation of the potential use of otolith chemistry in measuring population connectivity among Galápagos islands which are typical of the data sets used by otolith chemistry investigators (Ruttenberg & Warner 2006). For a portion of their study, these authors collected 13 benthic egg clutches of the damselfish *Stegastes beebei* from three areas in the western Galápagos for a total of 119 individuals ($n_1 = 50$, $n_2 = 45$, $n_3 = 24$). They then used laser-ablation inductively-coupled plasma mass spectrometry (ICPMS) to determine the concentration of 3 elements (^{86}Sr, ^{138}Ba, and ^{208}Pb, all normalized to calcium) in the natal sagittal otoliths. Their goal was to de-

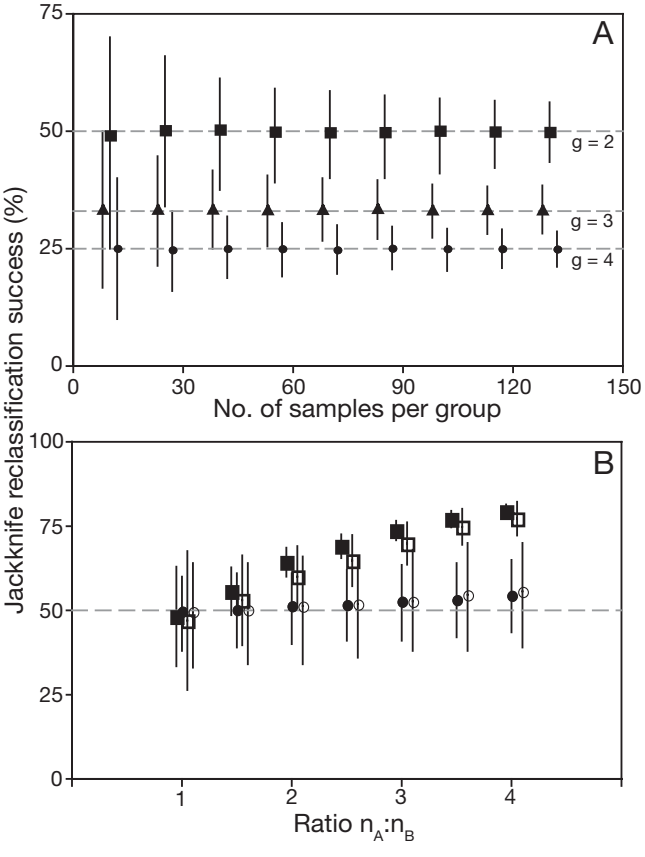

Fig. 1. Null expectations for linear DFA jackknife reclassification success (±95% confidence intervals) for randomly assembled data sets. Dashed lines: $1/g$ expectations for reclassification success; symbols offset for clarity. (A) Data sets with variable sample size, 2 (●), 3 (▲), or 4 (■) groups, and 3 predictor variables. (B) Data sets with 2 groups and uneven sample size; total sample sizes fixed at 50 (○,□) or 100 (●,■): n_A:n_B ratio of larger group sample size to smaller group sample size, reclassification success estimated with uniform priors (○,●) or priors proportional to sample sizes (□,■)

termine whether fish spawned in the 3 areas had distinct natal otolith chemical signatures that could be used to distinguish natal sites. Using quadratic DFA this analysis produced a 66.4% jackknife reclassification success (using uniform priors), significantly better than the 33.0% expected by chance ($p < 0.0001$).

As is often the case in this type of investigation, we are unable to propose informed prior estimates of population membership in this system, so we must classify unknowns using either quadratic DFA with uniform priors or ML. To compare the success rates of these methods we classified simulated data sets of 'unknown' individuals mixed in varying proportions. We used the sample means and covariance matrices in the original data as estimates of the true parameters for each area, then simulated natal signatures from each area as random draws from multivariate normal distributions with those parameters. We first generated a new 'known' training data set with sample sizes equal to those in the original sample (using the original sample as the training data set would have biased the classification of unknowns generated from the parameters of that sample). We then generated 6 sets of 500 'unknown' samples of 500 individuals: one set mixed in uniform proportions (167:167:166) and the other sets with arbitrarily skewed proportions ranging from 200:150:150 to 450:25:25. Because the true source population for each 'unknown' individual was actually known, we could determine whether classification was accurate. We recorded the fraction of successful classifications of the 'unknown' samples using the simulated training data set and either quadratic DFA with uniform prior probabilities or a maximum likelihood classification procedure. Maximum likelihood classifications were generated using Millar's expectation-maximization (EM) algorithm (Millar 1987); this algorithm is designed to calculate mixture proportions but we adapted it to also classify individuals to the group for which they have the highest likelihood.

With uniform priors, the DF correctly assigned a mean of 66.3% (SD = 2.1%) of the individuals in the uniformly mixed samples. This matches the jackknife reclassification success estimate of 66.4% almost exactly. Using ML classification did not improve accuracy for the uniformly mixed samples, but this method outperformed DFAs with uniform priors by an increasing margin as the mixing proportions of the samples became more skewed (Table 1).

Table 1. Results of classifying simulated 'unknown' samples (n = 500) using a training data set (n = 119) simulated from actual data and either quadratic DFA with uniform priors or maximum likelihood assignment using the EM algorithm. Each value is the mean (SD) of 500 simulations

Relative sample sizes (total = 500)	Mean percent classification accuracy	
	DFA (uniform priors)	Max. likelihood
167:167:166	66.3 (2.1)	65.9 (1.6)
200:150:150	66.2 (2.8)	67.6 (3.0)
300:100:100	68.0 (3.2)	72.8 (1.7)
350:75:75	70.4 (2.1)	78.1 (1.6)
400:50:50	71.6 (2.5)	84.4 (1.1)
450:25:25	72.9 (2.1)	91.8 (0.7)

DISCUSSION

Our simulations confirm that with equal sample sizes across groups, the null expectation of jackknife reclassification success is $1/g$. However, the estimated null reclassification values do change with unequal sample sizes, especially when priors are specified as proportional to relative sample sizes. When priors are specified in this way, the probability of correctly classifying individuals by chance alone increases greatly when 1 group dominates the data set. While this method may result in higher jackknife reclassification success values, the null expectation for reclassification success also increases, so these values are less likely to be significant. Furthermore, unless sampling effort for the training data set corresponds to the mixture of groups one expects to encounter in an unknown sample, jackknife values calculated in this way do not provide an informative estimate of reclassification success. When jackknife values are calculated for unbalanced data sets with uniform priors, the positive deviations from $1/g$ are still present but minimal, so $1/g$ appears to be a robust estimate of the null expectation. Regardless of the evenness of the sample, the variance of the null expectation increases sharply for small sample sizes, so more effort should be devoted to increasing sample size, even at the expense of sample evenness.

Given these results, it seems that except for the smallest data sets, most reasonably high jackknife values (calculated with uniform priors) will be statistically significant. Indeed, we suspect that most of the jackknife values reported in the 30 published DFA results from our literature search are statistically significant — the intuition of authors, editors, and reviewers that relatively high jackknife values are statistically meaningful is likely to be correct. However, it is important to note that low p-values do not always indicate a meaningful result; with a large data set or many groups, a poorly performing DF may have a jackknife value that is relatively low yet significantly greater than null expectations.

The classification of simulated Galápagos data yielded several lessons. First, the jackknife procedure does produce an unbiased estimate of reclassification success when the unknown sample is mixed with uniform proportions. This is not a new insight (Lachenbruch & Mickey 1968), but some investigators still report non-

jackknifed reclassification success, a practice which should be discouraged. More importantly, our results demonstrate the value of using ML procedures for classifying unknowns: >20% improvement in reclassification success over that achieved with DFA and uniform priors for some of our hypothetical examples. Given the simplicity and availability of numerical methods for estimating mixture proportions from the sample data (such as the EM algorithm used here), there seems to be little justification for further use of DFA with uninformative priors, even when knowledge of the relevant biology or oceanography is nebulous. In fact, ML methods similar to that used here may be a better choice for all classification tasks, except in cases where the distributional assumptions are severely violated (Millar 1990).

One factor affecting classification success that we did not directly address here is the number of predictor variables used. In general, additional data should improve the success of DFA and ML classification, as with any multivariate technique. For example, the jackknife statistic for the Galápagos data decreases from 66.3% to an average of 60.8% when only 2 of the 3 variables are used. However, Van Ness & Simpson (1976) found that additional variables must increase the distance between groups in multivariate space (i.e. add information) in order to improve reclassification success, while adding variables that are collinear with existing variables can hinder analysis (Williams 1983).

The continued interest among biologists in describing differences in fishery stocks and connectivity among fish populations and the availability of powerful statistical software is increasing the popularity of both otolith chemistry (Campana 2005) and DFA. In order to improve statistical rigor and accuracy using this analytical technique, we encourage investigators to use the randomization method presented here to assign p-values to jackknife reclassification success estimates and to use ML assignment when classifying unknown individuals. To this end, the Matlab code used to perform the tests described here is available at http://archive.lifesci.ucsb.edu/2006/07/17/01/dfa.zip.

Acknowledgements. We are grateful for the advice and guidance of S. Gaines, W. Rice, and R. R. Warner. J.W.W. was supported by a University of California Regents Fellowship. This is contribution 224 from the Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), funded primarily by the Gordon and Betty Moore Foundation and the David and Lucile Packard Foundation.

LITERATURE CITED

Cadrin SX (2000) Advances in morphometric identification of fishery stocks. Rev Fish Biol Fish 10:91–112

Campana SE (2005) Otolith science entering the 21st century. Mar Freshw Res 56:485–495

Campana SE, Chouinard GA, Hanson JM, Fréchet A (1999) Mixing and migration of overwintering Atlantic cod (*Gadus morhua*) stocks near the mouth of the Gulf of St. Lawrence. Can J Fish Aquat Sci 56:1873–1881

Campana SE, Chouinard GA, Hanson JM, Fréchet A, Brattey J (2000) Otolith elemental fingerprints as biological tracers of fish stocks. Fish Res 46:343–357

DeVries DA, Grimes CB, Prager MH (2002) Using otolith shape analysis to distinguish eastern Gulf of Mexico and Atlantic Ocean stocks of king mackerel. Fish Res 57:51–62

Gillanders BM (2002) Connectivity between juvenile and adult fish populations: do adults remain near their recruitment estuaries? Mar Ecol Prog Ser 240:215–223

Glick N (1978) Additive estimators for probabilities of correct classification. Pattern Recogn 10:211–222

Lachenbruch PA, Mickey MR (1968) Estimation of error rates in discriminant analysis. Technometrics 10:1–11

McGarigal K, Cushman S, Stafford S (2000) Multivariate statistics for wildlife and ecology research. Springer-Verlag, New York

Millar RB (1987) Maximum likelihood estimation of mixed stock fishery composition. Can J Fish Aquat Sci 44: 583–590

Millar RB (1990) Comparison of methods for estimating mixed stock fishery composition. Can J Fish Aquat Sci 47: 2235–2241

Ruttenberg BI, Warner RR (2006) Variation in the chemical composition of natal otoliths of a reef fish from the Galápagos Islands. Mar Ecol Prog Ser 328:225–236

Solow AR (1990) A randomization test for misclassification probability in discriminant analysis. Ecology 71:2379–2382

Thorrold SR, Jones CM, Swart PK, Targett TE (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

Thorrold SR, Latkoczy C, Swart PK, Jones CM (2001) Natal homing in a marine fish metapopulation. Science 291: 297–299

Van Ness JW, Simpson C (1976) On the effects of dimension in discriminant analysis. Technometrics 18:175–187

Wastell DG (1987) A simple randomisation procedure for validating discriminant analysis: a methodological note. Biol Psychol 24:123–127

Wells BK, Thorrold SR, Jones CM (2000) Geographic variation in trace element composition of juvenile weakfish scales. Trans Am Fish Soc 129:889–900

Wells BK, Thorrold SR, Jones CM (2003) Stability of elemental signatures in the scales of spawning weakfish, *Cynoscion regalis*. Can J Fish Aquat Sci 60:361–369

Williams BK (1983) Some observations of the use of discriminant analysis in ecology. Ecology 64:1283–1291

Editorial responsibility: Otto Kinne (Editor-in-Chief), Oldendorf/Luhe, Germany

Submitted: March 4, 2006; Accepted: May 24, 2006
Proofs received from author(s): December 29, 2006