The Pennsylvania State University

The Graduate School

Eberly College of Science

CLASSIFICATION OF PENNSYLVANIA BAT SPECIES BY DISCRIMINANT ANALYSIS OF ECHOLOCATION CALLS

A Thesis in

Statistics

by

Lauren E. Kraus

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

December 2012

The thesis of Lauren E. Kraus was reviewed and approved* by the following:

James L. Rosenberger
Professor of Statistics
Thesis Advisor

Durland Shumway
Assistant Professor of Statistics

David Hunter
Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

Indiana bats are an endangered species of bats found in the eastern half of the United States. As a result of White Nose Syndrome, the Indiana bat population nears extinction, making the location and identification of Indiana bats a priority. This thesis introduces the method of using the echolocation calls of Indiana bats and little brown bats to differentiate between the two species. First, the methods used to capture the ultrasonic calls made by bats are described. Next, the statistical method of discriminant function analysis is explained, and the various methods for evaluating classification results are explored. Finally, these methods are used on real bat echolocation data to discriminate between Indiana bats and little brown bats. Several statistical software packages and evaluation methods are used.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor, Dr. James L. Rosenberger, for his guidance and encouragement throughout the process of my research. Also, I would like to thank Amanda Brumbaugh for giving me invaluable insights into the world of bats and for providing the data on which this research is based.

Most importantly, I would like to thank my parents for their unwavering love and support, without which I would not be where I am today. Last, but certainly not least, I would like to thank my sisters, Erin and Kasey, and Meghan Kelly for their love and friendship.

1 Introduction to the Problem

1.1 The Indiana Bat

The Indiana bat, *Myotis sodalis*, is one of 27 subspecies of *Myotis* found in the United States (Sanborn, 1954). Indiana bats are small, weighing around a quarter of an ounce, have a wingspan of nine to eleven inches in flight, and have dark brown fur. They hibernate in caves and abandoned mines in the winter, and roost under the peeling bark of dead and dying trees in the summer. Indiana bats can be found over most of the eastern half of the United States, but almost half of the population hibernates in caves in southern Indiana (U.S. Fish and Wildlife Service, a).

While the Indiana bat population was once in the millions, it has steadily declined; the population was estimated to be 883,300 in the 1960s, and is estimated to be 424,708 today (U.S. Fish and Wildlife Service, b). In 1967, the Indiana bat was listed for protection under the Endangered Species Act. Reasons for the declining population include human disturbance during hibernation, commercialization of caves, loss of summer habitat, pesticides and other contaminants, and most recently white-nose syndrome (U.S. Fish and Wildlife Service, a). Despite the fact that white-nose syndrome has only recently started affecting this bat population, it is now the number one killer of Indiana bats.

1.2 White-Nose Syndrome

In February 2006, a caver photographed hibernating bats with a strange white substance on their muzzles in a cave 40 miles west of Albany, N.Y. During the following winter, while conducting a routine census of bats in several caves in the same area, biologists with the New York Department of Environmental Conservation discovered bats behaving strangely, bats with white noses, and a few hundred dead bats. They

named the affliction "White-Nose Syndrome" (WNS) for the white fungus that appears on the muzzle and other body parts of hibernating bats (U.S. Fish and Wildlife Service, d).

So far, nine species of bats have been affected by WNS: big brown bat (*Eptesicus fuscus*), eastern small-footed bat (*Myotis leibii*), Indiana bat (*Myotis sodalis*), little brown bat (*Myotis lucifugus*), northern long-eared bat (*Myotis septentrionalis*), tricolored bat (*Perimyotis subflavus*), southeastern bat (*Myotis austroriparius*), cave bat (*Myotis velifer*), and gray bat (*Myotis grisescens*). Of these nine species, the Indiana bat and the gray bat are the only species currently listed as endangered (Foley et al., 2011). Since its origin in New York, at the end of the 2011-2012 winter, bats with WNS have been confirmed in 19 states and four Canadian provinces: Alabama, Connecticut, Delaware, Indiana, Kentucky, Maine, Maryland, Massachusetts, Missouri, New Hampshire, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Tennessee, Vermont, Virginia, West Virginia, New Brunswick, Nova Scotia, Ontario, and Quebec. A map showing the geographic spread of WNS can be seen in Figure 1.1.

As a result of the recent detection of WNS among bat populations, much about the disease remains a mystery. Despite the many questions surrounding the causes of WNS, two important distinctions have been made by researchers. First, all of the bats affected by WNS have white fungus on their bodies. The fungus associated with WNS is *Geomyces destructans*, a cold-loving fungus that thrives in the dark at low temperatures and high humidity. Researchers do not know if the fungus is the direct cause of the disease, or merely a symptom. It is believed that the fungus came from Europe and is new to North America. If the fungus is indeed the cause of WNS, the fact that the fungus is not indigenous to North America could explain why bat species in the United States and Canada show little to no resistance to WNS (Vories,

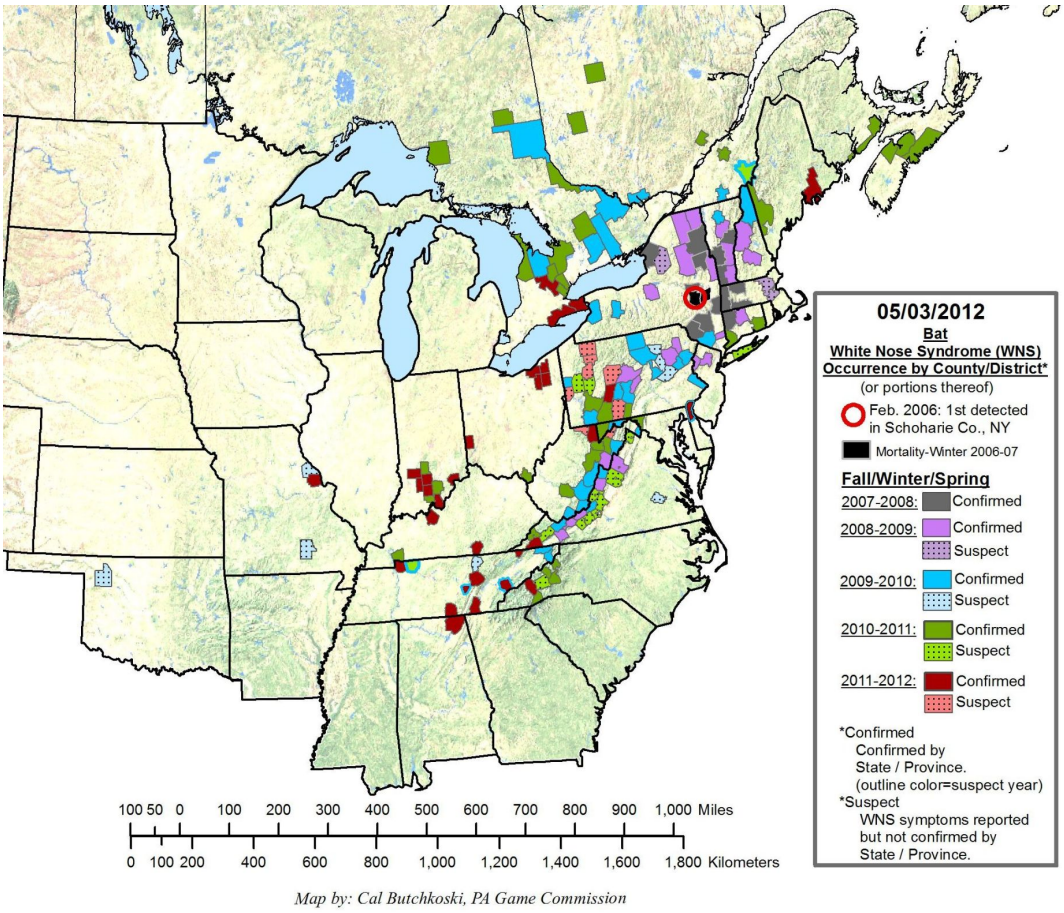

Figure 1.1: White-nose syndrome geographic distribution as of May 3, 2012 (U.S. Fish and Wildlife Service, d).

2010). The second fact about WNS is that it affects hibernating bats. Bats affected by WNS seem to deplete their fat reserves before the end of the winter, causing them to awake from hibernation and hunt for food. During the winter, few insects can be found, so the bats starve to death. The exact cause of the bats' depleted reserves is unknown, but the two main theories are that the bats are not finding enough food to build large enough fat reserves for winter, or that they are burning through their reserves too quickly (Cohn, 2008).

In spite of the fact that the mystery of WNS is yet to be solved, what is known, is that nearly all bats with WNS have died (Cohn, 2008). WNS has killed more than

5.5 million bats in the Northeast and Canada, and in some hibernation areas, 90 to 100 percent of the bats have died (U.S. Fish and Wildlife Service, d). It is feared that if WNS continues to spread at the same rapid pace across North America, more species will become endangered and some may become extinct. The location and identification of Indiana bats in the wild has become of upmost importance because of the endangered status of the species and its rapidly declining population. Once existing Indiana bat communities are located, the hope is that through protection of their habitat, extinction of the species will be avoided.

2 Acoustic Identification of Bats

When studying bats, there are two main techniques used to identify and inventory bats: capture techniques and acoustic techniques. In this study, the focus is on acoustic techniques. These techniques involve the use of bat detectors to capture the ultrasonic calls made by bats, followed by an analysis of the call signal. There are many advantages to identification using acoustic techniques. Bat detectors are non-intrusive, can be used repeatedly at one site, and can be used in areas where capture techniques are difficult to perform (Kuenzi and Morrison, 1998). It has also been shown that acoustic methods are able to detect more bats in an area than capturing techniques (O'Farrell and Gannon, 1999). However, there are some disadvantages to using ultrasonic devices. Recognition of species can be complicated because of the geographic and individual variation in echolocation calls within each species (Kuenzi and Morrison, 1998), and it has been shown that not all species are detected equally by acoustic detectors (Gannon and Sherwin, 2000).

There are three steps to collecting acoustic information: an ultrasonic call is transmitted through the air, the call is captured by the microphone of a bat detector, and the call is transformed into audible sound (Pettersson, 2000). There are several different methods used to transform the ultrasonic signals into audible signals; the most popular are the heterodyne, frequency division, zero-cross, and time expansion techniques.

2.1 The SonoBat Bat Detector

For this study, the calls are captured using a SonoBat detector, which uses the time-expansion method. With this method, a high frequency signal (the bat call) is recorded and then played back at a slower speed. As a result, the new signal is

longer than the original and the frequency is lower. The detector stores the original signal, so that the transformed signal contains the same information as the original signal. The advantage of this transformation technique is that since the signals are stretched out in time, it is often possible to hear details in the signal that would not otherwise be apparent if the signal were listened to in real time (Pettersson, 2000). Once bat echolocation calls are captured by a SonoBat detector, 76 parameters are measured for each call. A list of these parameters, along with a description of what each parameter measures, can be found in Table A.1 in the Appendix.

The SonoBat detector has advantages over other bat detectors. Fenton (2000) showed that a bat detector using zero-cross techniques (the AnaBat II Bat Detector) was significantly less sensitive to bat echolocation calls than a SonoBat detector, meaning that the SonoBat detector was able to detect more bats. Also, Szewczak (2000) showed that calls from *Myotis* species, such as those from *Myotis sodalis* (the Indiana bat), were more easily discriminated from each other using the SonoBat detector.

2.2 Analysis of Bat Echolocation Calls

The main quantitative method of analysis used to identify echolocation calls of bat species is discriminant function analysis (DFA). This method has been used to varying levels of success in many studies. Parsons and Jones (2000) used quadratic DFA to identify 14 species of bats in Britain. They used 13 parameters in their analysis, and had an overall correct classification rate of 81%. Russo and Jones (2002) used quadratic DFA to identify 18 species of bats in Italy. Using 6 parameters in their analysis, they achieved an overall correct classification rate of 82%, with the correct classification rates for each species ranging from 38% to 98%. Preatoni et al. (2005) used quadratic DFA to identify 7 species of *Myotis* in Italy. They used 7 parameters

in their analysis, which resulted in an overall correct classification rate of 91% and correct classification rates for each species ranging from 70% to 100%. Britzke et al. (2011) used five different types of DFA to identify 12 species of bats in the eastern United States. With the five types of DFA, they used 10 parameters and had overall correct classification rates ranging from 87% to 93%. The most successful analysis used a mixture discriminant analysis model based on adaptive regression splines, which had correct classification rates for each species ranging from 83% to 100%.

In their 2002 study, Britzke et al. (2002) looked at four species of bats in the eastern United States. Their study had two goals: to classify the calls into species, and to determine the likelihood of the absence of a particular species in a sampled community. The researchers collected search-phase echolocation calls of eight species of bat, but they focused their analysis on the four species of *Myotis*: gray bat, little brown bat, northern bat, and Indiana bat. They recorded 552 call sequences using the Anabat detector, and they used Analook to determine values for 10 parameters of each call sequence. A linear DFA on these parameters, followed by cross-validation was performed, and they achieved rates of accuracy ranging from 93% to 100%. Britzke et al. then used these rates to construct a likelihood-ratio test of the null hypothesis that a species was absent from the community.

Britzke et al. (2002) begin constructing their likelihood-ratio test by stating that they are operating under the assumptions that the correct classification rates are estimated precisely and are uniform across communities, each sequence identified by the model is independent of other identifications, and that the bat species mentioned in the study are the only species of bat potentially present in a community. When building their test, they included all eight species of bats, not just the *Myotis* species. The null hypothesis that a species is absent from the sampled community is then tested using a generalized likelihood-ratio test. Let ϕ_{ij} be the conditional probability

that an individual of species j is identified as species i. Let $\hat{\theta}_i$ represent the proportion of species i in the sampled community, meaning that it is the probability that a sampled bat actually belongs to species i. The probabilities $\boldsymbol{\theta}$ are based on the population of bat communities, and are therefore unknown. Because of this, they instead work with the empirical probabilities $\hat{\boldsymbol{\theta}}$. Therefore, they are testing the null hypothesis that $\theta_k = 0$ for a particular species k.

The probability that a sampled bat is identified as species i is:

$$p_i = \sum_{j=1}^{8} \phi_{ij} \hat{\theta}_i$$

If a sample of N bats from the community yields n_i identifications of species i, then the likelihood function is defined as:

$$L(\boldsymbol{n}, \hat{\boldsymbol{\theta}}) = \binom{N}{\boldsymbol{n}} \prod_i \left(\sum_{j=1}^{8} \phi_{ij} \hat{\theta}_j \right)^{n_i} = \binom{N}{\boldsymbol{n}} \prod_i p_i^{n_i}$$

A test of $H_0 : \theta_k = 0$ versus $H_1 : \theta_k > 0$ can the be performed using the generalized likelihood ratio:

$$\lambda = \frac{L_{max}(\boldsymbol{n}, \hat{\boldsymbol{\theta}}, \hat{\theta}_k = 0)}{L_{max}(\boldsymbol{n}, \hat{\boldsymbol{\theta}})}$$

The denominator of λ is the likelihood function evaluated at the maximum likelihood estimates of $\hat{\boldsymbol{\theta}}$, with the constraint that all $\hat{\boldsymbol{\theta}}$ are non-negative and sum to one, while the numerator is the likelihood function evaluated at the maximum likelihood estimates of $\hat{\boldsymbol{\theta}}$, with the additional constraint that $\hat{\theta}_k = 0$. If the null hypothesis is true, then $-2 \log \lambda$ follows a chi-squared distribution with 1 degree of freedom for large N.

3 Classification by Discriminant Analysis

The goal of this study is to develop a method for detecting the presence of Indiana bats in a sampled community. When analyzing bat echolocation calls, this analysis will follow the requirements stated in the "Rangewide Summer Guidance IBAT Survey Protocol" created by the U.S. Fish and Wildlife Service (U.S. Fish and Wildlife Service, c). The following are the requirements for analysis of recorded echolocation calls, stated by the protocol:

1. Any call identification analysis program should be based on a large call library.
2. The analysis process is quantitative, preferably automated.
3. The process should include filtering to remove extraneous noise and no-bat files.
4. The process should include an "unknown" category for classifying calls that are not characteristic of species in the call library to ensure that such calls are not forced to species identification.
5. Accuracy rates should be derived through cross-validation.
6. All analysis programs should utilize maximum likelihood to determine species presence, rather than relying on a single call sequence.

The ideas and methods behind discriminant analysis go back to 1920 and the English statistician Karl Pearson. Pearson proposed what was called the coefficient of racial likeness (CRL), which is a type of intergroup distance index. While the CRL was studied by G.M. Morant in the 1920s, another distance index was being created in India. This index was formalized by P.C. Mahalanobis in the 1930s. In the 1930s, R.A. Fisher translated the idea of multivariable intergroup distance to that of a linear combination of variables derived for the purpose of intergroup discrimination. The

idea of discriminant analysis was formerly brought forward in 1936, with Fisher's paper "The use of multiple measurements in taxonomic problems", which appeared in *Annals of Eugenics*. Since then, many extensions and improvements of Fisher's ideas have been made (Huberty, 1994).

The term "discriminant analysis" encompasses two distinct multivariate procedures: discrimination and classification. Discrimination deals with separating distinct sets of objects, while classification deals with allocating new objects to previously defined groups. Discrimination is often used in order to investigate observed differences when causal relationships are not well understood; classification methods give well-defined rules, which can be used to assign new objects to groups. The differences between these two procedures can be seen when comparing their goals. The goal of discrimination is to describe the differential features of objects/observations from several known populations. This method tries to find "discriminants" whose numerical values are such that the populations are separated as much as possible. The goal of classification is to sort objects/observations into two or more classes. The emphasis of this method is on deriving a rule that can be used to assign new objects/observations to the classes (Johnson and Wichern, 2007).

Together, these two procedures, discrimination and classification, are used in the two different types of discriminant analysis: predictive discriminant analysis (PDA) and descriptive discriminant analysis (DDA). PDA focuses on the prediction of group membership; with this method, the goal is to assign a new object to a group. In this setting, the multiple response variables are the predictor variables and the grouping variable is the outcome. Using discriminant functions to interpret effects revealed via a multivariate analysis of variance (MANOVA) is DDA. With DDA, the multiple response variables are the outcome (dependent) variables and the grouping variable(s) is(are) the predictor(s). Specifically, DDA is a study of the effects that the outcome

variables have on group separation (Huberty, 1994). For the purposes of this study, the focus will be on PDA.

To start the process of PDA, suppose there are samples from k populations of size n_g, $g = 1, 2, ..., k$, with p measures of each of the $N = \sum_g n_g$. Using the $N \times p$ data matrix, the goal of PDA is to determine from which of the k populations an $(N+1)$st unit is most likely to have been randomly sampled. To do this, the information in the $N \times p$ data matrix is used to set up a rule for making the assignment. In PDA, it is assumed that the k groups of n_g units represent k meaningful populations and that any unit to be classified does in fact belong to one of the k populations.

3.1 Distance Measures

Before getting into the mechanics of classification, a short discussion on measures of distance is required. There are three types of distances: (1) unit to unit, (2) centroid to centroid, and (3) unit to centroid. Unit to unit distance will be discussed first. When discussing distance, it is simplest to start with the distance, d_{AB}, between two points in bivariate space, $A : (X_{1A}, X_{2A})$ and $B : (X_{1B}, X_{2B})$. By the Pythagorean Theorem, the Euclidean distance is derived:

$$d_{AB}^2 = (X_{1A} - X_{1B})^2 + (X_{2A} - X_{2B})^2 = \sum_{i=1}^{2} (X_{iA} - X_{iB})^2 = [X_{\boldsymbol{A}} - X_{\boldsymbol{B}}]'[X_{\boldsymbol{A}} - X_{\boldsymbol{B}}]$$

where X_A and X_B are 2×1 column vectors. This distance index is appropriate if the measures on X_A and X_B are uncorrelated and $\sigma_A^2 = \sigma_B^2 = 1$. Extending Euclidean distance to a general p-variate space:

$$d_{AB}^2 = \sum_{i=1}^{p} (X_{iA} - X_{iB})^2 = [X_{\boldsymbol{A}} - X_{\boldsymbol{B}}]'[X_{\boldsymbol{A}} - X_{\boldsymbol{B}}]$$

where X_A and X_B are $p \times 1$ column vectors. Similar to the bivariate case, for the p-variate index, it is assumed that there are uncorrelated variables and that the $p \times p$ covariance matrix, Σ, is an identity matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1p}\sigma_{1p} \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2p}\sigma_1\sigma_p \\ \vdots & & \ddots & \vdots \\ \rho_{p1}\sigma_p\sigma_1 & \rho_{p2}\sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Since it is rarely the case where there are uncorrelated variables, the intercorrelation of variables must be taken into account when measuring distance. Therefore, a generalized distance index, Δ_{AB}, attributed to P.C. Mahalanobis, which takes both unequal variance and nonzero intercorrelations into account, is used. The generalized (squared) distance between point A, defined by column vector $\boldsymbol{X}_A$, and point B, defined by column vector $\boldsymbol{X}_B$, is:

$$\Delta_{AB}^2 = [X_{\boldsymbol{A}} - X_{\boldsymbol{B}}]'\Sigma^{-1}[X_{\boldsymbol{A}} - X_{\boldsymbol{B}}] \tag{3.1}$$

where Σ is the population covariance matrix.

The second index of distance is a measure of the distance between two points where each point represents a vector of means on the p variables. A vector of means is called a centroid, and the centroid for a population g is denoted by $\boldsymbol{\mu}_g' = [\mu_{1g}, \mu_{2g}, ..., \mu_{pg}]$, where μ_{ig} is the mean of variable i in population g. The distance between two centroids, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is then defined as:

$$\Delta_{12} = \left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^{1/2} \tag{3.2}$$

where Σ is the covariance matrix common to the two populations.

The third index of distance is used when one point represents a vector of p observations on an experimental unit and the other point represents a centroid for a population. The distance between $\boldsymbol{X}_u$, the observation vector for unit u, and $\boldsymbol{\mu}_g$, the centroid for population g, is

$$\Delta_{ug} = \left[(\boldsymbol{X}_u - \boldsymbol{\mu}_g)'\Sigma_g^{-1}(\boldsymbol{X}_u - \boldsymbol{\mu}_g)\right]^{1/2} \tag{3.3}$$

where Σ_g is the covariance matrix for population g. This measure of distance is important for classification analyses, where the goal is to classify a unit into the population to which the unit is nearest. That is, unit u is classified into population g if Δ_{ug} is smaller than $\Delta_{ug'}$ for all $g' \neq g$ and g, $g' = 1, 2, ..., k$. To summarize, there are three types of distance: unit to unit, defined in (3.1), centroid to centroid, defined in (3.2), and unit to centroid, defined in (3.3) (Huberty, 1994).

3.2 Classification Rules

A classification rule in discriminant analysis is based on the maximum likelihood principle; an object is assigned to the population in which its observation vector has the greatest likelihood of occurrence. Again, start with k populations. It must be assumed that all k populations have the same density function. For example, it can be assumed that all density functions are multivariate normal. Let f represent this common density function. The maximum likelihood rule assigns unit u to population g if the likelihood of the observation vector, X_u is greater for population g than for any other group. Formally, this rule is stated as:

$$\text{Assign unit } u \text{ to population } g \text{ if } f(\boldsymbol{X}_u|g) > f(\boldsymbol{X}_u|g') \text{ for } g' \neq g \tag{3.4}$$

The maximum likelihood rule may also be stated in terms of probabilities. The first is in terms of *inverse probabilities* $P(\boldsymbol{X}|g)$, where $P(\boldsymbol{X}|g)$ is the probability that a randomly selected unit has a profile close to $\boldsymbol{X}$ given that the unit is a member of population g. $P(\boldsymbol{X}|g)$ is proportional to $f(\boldsymbol{X}|g)$ in its limit, so the maximum likelihood rule can also be stated as:

$$\text{Assign unit } u \text{ to population } g \text{ if } P(\boldsymbol{X}_u|g) > P(\boldsymbol{X}_u|g') \text{ for } g' \neq g \tag{3.5}$$

Lastly, the maximum likelihood rule can be stated in terms of *posterior probabilities* $P(g|\boldsymbol{X}_u)$, where $P(g|\boldsymbol{X}_u)$ is the probability of unit u belonging to group g given that the unit has observation vector $\boldsymbol{X}_u$. Note that:

$$P(g|\boldsymbol{X}_u) = \frac{P(\boldsymbol{X}_u|g)}{\sum_{g'=1}^{k} P(\boldsymbol{X}_u|g')} \tag{3.6}$$

Therefore, the maximum likelihood rule can also be stated as:

$$\text{Assign unit } u \text{ to population } g \text{ if } P(g|\boldsymbol{X}_u) > P(g'|\boldsymbol{X}_u) \text{ for } g' \neq g \tag{3.7}$$

where $P(g|\boldsymbol{X}_u)$ is defined as in (3.6). Regardless of the choice of which version of the maximum likelihood to use, k values need to be estimated for each unit: k values of $f(\boldsymbol{X}_u|g)$ for (3.4), k values of $P(\boldsymbol{X}_u|g)$ for (3.5), and k values of $P(g|\boldsymbol{X}_u)$ for (3.7).

Looking at the three versions of the maximum likelihood rule, particularly the last two, it is clear that the quality of the rule depends on the quality of the estimates of the probabilities $P(\boldsymbol{X}_u|g)$. The quality of these probabilities depends on the size and representativeness of the k original samples on which the estimates are based. Because of this, it is often a good idea to take the relative sizes of the populations into

account. Let π_g denote the proportion of units in the total universe (the combination of all k populations) that is in population g. This means that if a unit is randomly selected from the universe, the probability that it belongs to population g is equal to π_g. π_g is known as the *prior probability* of membership in population g. These prior probabilities can be taken into consideration when calculating values of $P(g|\boldsymbol{X}_u)$ for (3.7). The product $\pi_g P(g|\boldsymbol{X}_u)$ denotes the joint probability that a randomly selected unit belongs to population g and at the same time has a profile close to $\boldsymbol{X}$, and it can then be used to calculate values of $P(\boldsymbol{X}_u|g)$ using Bayesian methodology. The posterior probability of unit u belonging to population g, given profile vector $\boldsymbol{X}$ is:

$$P(g|\boldsymbol{X}_u) = \frac{\pi_g P(\boldsymbol{X}_u|g)}{\sum_{g'=1}^{k} \pi_{g'} P(\boldsymbol{X}_u|g')} \tag{3.8}$$

The maximum Bayesian probability rule is then:

$$\text{Assign unit } u \text{ to population } g \text{ if } P(g|\boldsymbol{X}_u) > P(g'|\boldsymbol{X}_u) \text{ for } g' \neq g \tag{3.9}$$

where $P(g|\boldsymbol{X}_u)$ is defined as in (3.8).

As with the previous three versions of the rule, k values of $P(g|\boldsymbol{X}_u)$ need to be calculated for each unit. Since the value in the denominator of (3.8) is the same for all populations, it is possible to have the rule be based solely on the k values of $\pi_g P(\boldsymbol{X}_u|g)$. In addition, since the $P(\boldsymbol{X}_u|g)$ values are proportional to $f(\boldsymbol{X}_u|g)$, the rule can be based on the k value of $\pi_g f(\boldsymbol{X}_u|g)$. Rao calls the values of $\pi_g f(\boldsymbol{X}_u|g)$

discriminant scores (Rao, 1973). (3.8) may be stated equivalently as

$$P(g|\boldsymbol{X}_u) = \frac{\pi_g f(\boldsymbol{X}_u|g)}{\sum_{g'=1}^{k} \pi_{g'} f(\boldsymbol{X}_u|g')} \tag{3.10}$$

When choosing which version of the classification rule to use, there are several factors to take into account. Intuitively, one wants to use a classification rule that minimizes the misclassification error. However, there are different ways to do this. Using the conditional Bayesian posterior probabilities, as in (3.8), minimizes the total number of misclassification errors. Using the maximum likelihood rule (3.7) minimizes the total proportion of misclassification errors (Huberty, 1994). Lastly, it is possible to refine the Bayes procedure to minimize the total cost of misclassification errors.

To discuss misclassification costs, examine the example of $k = 2$ populations. In this case, π_1 is the prior probability of membership in population 1 and π_2 is the prior probability of membership in population 2. In this case, there are the following four probabilities:

$$\begin{aligned}
P(\text{object is correctly classified as 1}) &= P(\text{object comes from 1 and is classified as 1}) \\
&= P(1|1)\pi_1 \\
P(\text{object is misclassified as 1}) &= P(\text{object comes from 2 and is classified as 1}) \\
&= P(1|2)\pi_2 \\
P(\text{object is correctly classified as 2}) &= P(\text{object comes from 2 and is classified as 2}) \\
&= P(2|2)\pi_2 \\
P(\text{object is misclassified as 2}) &= P(\text{object comes from 1 and is classified as 2}) \\
&= P(2|1)\pi_1
\end{aligned}$$

The cost of misclassification can then be defined by a cost matrix:

		Classified as	
		1	2
True Population	1	0	$c(2\|1)$
	2	$c(1\|2)$	0

Looking at this matrix, it can be seen that that the cost of misclassification is zero if the correct classification is made, $c(1|2)$ if an observation from population 2 is classified as belonging to population 1, and $c(2|1)$ if an observation from population 1 is classified as belonging to population 2. For any classification rule, the *expected cost of misclassification* (ECM) is:

$$\text{ECM} = c(2|1)P(2|1)\pi_1 + c(1|2)P(1|2)\pi_2 \tag{3.11}$$

This value is calculated by multiplying the entries in the cost matrix by their probabilities of occurrence, seen on the previous page. Next, the generalization of this formula for the case where there are k populations, $k > 2$, will be found.

The conditional expected cost of misclassifying an observation from population 1 into populations 2, or 3,..., or k is:

$$\text{ECM}(1) = P(2|1)c(2|1)+P(3|1)c(3|1)+...+P(k|1)c(k|1) = \sum_{g'=2}^{k} P(g'|1)c(g'|1) \tag{3.12}$$

In the same way, the expected costs of misclassification ECM(2),...,ECM(k) can be found. The overall ECM is then calculated as:

$$\text{ECM} = \pi_1\text{ECM}(1) + \ldots + \pi_k\text{ECM}(k) = \sum_{g=1}^{k} \pi_g \left(\sum_{\substack{g'=1 \\ g' \neq g}}^{k} P(g'|g)c(g'|g) \right) \tag{3.13}$$

Using this, a classification rule that minimizes ECM is created. This rule is:

$$\text{Assign unit } u \text{ to population } g \text{ if } \sum_{g=1}^{k} \pi_g \left(\sum_{\substack{g'=1 \\ g' \neq g}}^{k} P(g'|g) c(g'|g) \right) \text{ is a minimum} \quad (3.14)$$

(Johnson and Wichern, 2007).

3.3 Classification Rules for Multivariate Normal Populations

As previously stated, the maximum probability rule involving posterior probabilities of group membership, the probabilities defined in (3.8) or (3.10), will minimize the total number of misclassification errors. This rule can only be used if the probability density functions, $f(\boldsymbol{X}|g)$, are known, which is only the case when the distribution parameters, Σ and $\boldsymbol{\mu}$ are known. Since the distribution parameters are rarely known, classification rules that use estimates of the parameters must be constructed, and therefore use estimates of the density values, $\hat{f}(\boldsymbol{X}_u|g)$. The most common approach to this problem is to assume the data fit a multivariate normal model, estimate the model parameters using the data, and construct a rule using these estimates.

The p-variate normal probability density function is defined as:

$$f(\boldsymbol{X}|g) = \frac{1}{\sqrt{(2\pi)^p}\sqrt{|\Sigma_g|}} \exp\left[-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_g)'\Sigma_g^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_g)\right] \quad (3.15)$$

where $|\Sigma_g|$ is called the generalized variance of the set of p variables. For $p > 2$, calculating $f(\boldsymbol{X}|g)$ can be very difficult, since calculating the value of the determinant $|\sigma_g|$ and the value of the quadratic form

$$(\boldsymbol{X} - \boldsymbol{\mu}_g)'\Sigma_g^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_g) \quad (3.16)$$

is difficult. The quadratic form (3.16) is a distance index of the type seen in (3.3), therefore the square of the distance from $\boldsymbol{X}_u$, the observation vector from unit u, to the centroid of the population g, $\boldsymbol{\mu}_g$, is $\Delta^2_{ug} = (\boldsymbol{X}_u - \boldsymbol{\mu}_g)'\Sigma_g^{-1}(\boldsymbol{X}_u - \boldsymbol{\mu}_g)$. The estimate of the multivariate normal probability density function (3.15) is then

$$\hat{f}(\boldsymbol{X}|g) = \frac{1}{\sqrt{(2\pi)^p}\sqrt{|S_g|}} \exp\left[-\frac{1}{2}(\boldsymbol{X} - \bar{\boldsymbol{X}}_g)'S_g^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}}_g)\right] \tag{3.17}$$

where $\bar{\boldsymbol{X}}_g$ is the $p \times 1$ vector of means for sample g, and S_g is the $p \times p$ covariance matrix for sample g. The main diagonal elements of S_g are the p variance defined as:

$$s_g^{(i)^2} = \frac{1}{n_g - 1}\sum_u \left(X^{(i)} - \bar{X}_g^{(i)}\right)^2$$

and the off-diagonal elements are covariances defined as:

$$s_g^{(i,i')} = \frac{1}{n_g - 1}\sum_u \left(X^{(i)} - \bar{X}_g^{(i)}\right)\left(X^{(i')} - \bar{X}_g^{(i')}\right)$$

The sample Mahalanobis distance index for the squared distance between an observation vector unit u and the centroid for sample g is then defined as

$$D^2_{ug} = (\boldsymbol{X}_u - \bar{\boldsymbol{X}}_g)'S_g^{-1}(\boldsymbol{X}_u - \bar{\boldsymbol{X}}_g) \tag{3.18}$$

Expression (3.17) can then be stated for unit u as:

$$\hat{f}(\boldsymbol{X}_u|g) = (2\pi)^{-p/2}|S_g|^{-1/2}\exp\left(-\frac{1}{2}D^2_{ug}\right) \tag{3.19}$$

When forming the classification rules based on normality, the posterior probabilities as seen in (3.10) are used. Replacing the parameters with their estimates, the

result is:

$$\hat{P}(g|\boldsymbol{X}_u) = \frac{q_g \hat{f}(\boldsymbol{X}_u|g)}{\sum_{g'=1}^{k} q_{g'} \hat{f}(\boldsymbol{X}_u|g')} \tag{3.20}$$

where $q_g = \hat{\pi}_g$. Substituting (3.19) into the above formula, the resulting estimate is:

$$\hat{P}(g|\boldsymbol{X}_u) = \frac{q_g |S_g|^{-1/2} \exp\left(-\frac{1}{2} D^2_{ug}\right)}{\sum_{g'=1}^{k} q_{g'} |S_{g'}|^{-1/2} \exp\left(-\frac{1}{2} D^2_{ug'}\right)} \tag{3.21}$$

Therefore, the maximum probability rule for the p-variate normal case is:

$$\text{Assign unit } u \text{ to population } g \text{ if } \hat{P}(g|\boldsymbol{X}_u) > \hat{P}(g'|\boldsymbol{X}_u) \text{ for } g' \neq g \tag{3.22}$$

where $\hat{P}(g|\boldsymbol{X}_u)$ is defined as in (3.21). As mentioned earlier, the density values $\hat{f}(\boldsymbol{X}|g)$ are proportional to the probability values $\hat{P}(\boldsymbol{X}|g)$. Therefore, the posterior probability estimates defined in (3.20) may be written as

$$\hat{P}(g|\boldsymbol{X}_u) = \frac{q_g \hat{P}(\boldsymbol{X}_u|g)}{\sum_{g'=1}^{k} q_{g'} \hat{P}(\boldsymbol{X}_u|g')} \tag{3.23}$$

For a given unit, the value of the denominator in (3.23) is the same for all groups. Therefore, the denominator may be ignored and and the maximum probability rule seen in (3.22) may be written as

$$\text{Assign unit } u \text{ to population } g \text{ if } q_g \hat{P}(\boldsymbol{X}_u|g) > q_{g'} \hat{P}(\boldsymbol{X}_u|g') \text{ for } g' \neq g \tag{3.24}$$

When looking at the classification rule stated in (3.22), the special case where

it is assumed that the k population covariance matrices are equal, meaning that $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k = \Sigma$, must be considered. When this is the case, then the estimator for Σ is the pooled sample $p \times p$ covariance matrix, S. The elements of S along the main diagonal are the p pooled sample variances seen in univariate analyses of variance. Specifically, the ith main diagonal element of S is the pooled sum of squares for variable i divided by $N - k$. The (i, i') off-diagonal element is the pooled sum of products for variables i and i' divided by $N - k$. The squared distance of unit u from the centroid of sample g then becomes a special case of (3.18), and is defined as:

$$D^{*2}_{ug} = (\boldsymbol{X}_u - \bar{\boldsymbol{X}}_g)' S^{-1} (\boldsymbol{X}_u - \bar{\boldsymbol{X}}_g) \tag{3.25}$$

Then, the estimated likelihood seen in (3.19) becomes

$$\hat{f}(\boldsymbol{X}|g) = (2\pi)^{-p/2} |S|^{-1/2} \exp\left(-\frac{1}{2} D^{*2}_{ug}\right) \tag{3.26}$$

Since $(2\pi)^{-p/2}|S|^{-1/2}$ is common to the numerator and the denominator of (3.20), for this special case, $\hat{P}(g|\boldsymbol{X}_u)$ is:

$$\hat{P}(g|\boldsymbol{X}_u) = \frac{q_g \exp\left(-\frac{1}{2} D^{*2}_{ug}\right)}{\sum_{g'=1}^{k} q_{g'} \exp\left(-\frac{1}{2} D^{*2}_{ug'}\right)} \tag{3.27}$$

The maximum probability rule for the p-variate normal case where the k population covariance matrices are equal is then

$$\text{Assign unit } u \text{ to population } g \text{ if } \hat{P}(g|\boldsymbol{X}_u) > \hat{P}(g'|\boldsymbol{X}_u) \text{ for } g' \neq g \tag{3.28}$$

where $\hat{P}(g|\boldsymbol{X}_u)$ is defined as in (3.27) (Huberty, 1994).

3.4 Classification Functions

As previously stated, the form of the maximum probability rule stated in (3.22) may be stated in terms of only the numerators in (3.21). This means that the rule (3.22) may be equivalently stated in terms of maximizing $q_g|S_g|^{-1/2}\exp\left(-\frac{1}{2}D^2_{ug}\right)$. This is equivalent to maximizing the natural logarithm of the following product:

$$Q_{ug} = \log q_g - \frac{1}{2}\log|S_g| - \frac{1}{2}D^2_{ug} \tag{3.29}$$

Therefore, the maximum probability rule for the p-variate normal case may also be written as:

$$\text{Assign unit } u \text{ to population } g \text{ if } Q_{ug} > Q_{ug'} \text{ for } g' \neq g \tag{3.30}$$

where Q_{ug} is defined as in (3.29). It can be shown that Q_{ug} is quadratic in $\boldsymbol{X}_u$, and is therefore called a *quadratic classification function* (QCF). The *quadratic classification rule* therefore involves calculating k QCF values of each unit u, and then u is assigned to the population whose sample yields the largest QCF value.

Return now to the special case where the population covariance matrices are equal and S is the estimator of the common matrix. In this setting, maximizing $\hat{P}(g|\boldsymbol{X}_u)$, defined as in (3.27), is equivalent to maximizing $q_g\exp\left(-\frac{1}{2}D^{*2}_{ug}\right)$. This is equivalent to maximizing the natural logarithm of the following product:

$$\log q_g - \frac{1}{2}D^{*2}_{ug} = \log q_g - \frac{1}{2}(\boldsymbol{X}_u - \bar{\boldsymbol{X}}_g)'S^{-1}(\boldsymbol{X}_u - \bar{\boldsymbol{X}}_g) \tag{3.31}$$

After some matrix algebra, the term $-\frac{1}{2}\boldsymbol{X}'_u S^{-1}\boldsymbol{X}_u$ can be calculated. This term is common for all g for a given unit u, therefore, this term may be ignored for classifi-

cation purposes. As a result, maximizing (3.31) is equivalent to maximizing

$$L_{ug} = (\bar{\boldsymbol{X}}'_g S^{-1})\boldsymbol{X}_u - \frac{1}{2}\bar{\boldsymbol{X}}'_g S^{-1}\bar{\boldsymbol{X}}_g + \log q_g \tag{3.32}$$

Therefore, the maximum probability rule for the p-variate normal case may also be written as:

$$\text{Assign unit } u \text{ to population } g \text{ if } L_{ug} > L_{ug'} \text{ for } g' \neq g \tag{3.33}$$

where L_{ug} is defined as in (3.32). L_{ug} is linear in $\mathbf{X}_u$, and is therefore called a *linear classification function* (LCF), making the rule in (3.33) a *linear classification rule*.

Looking at (3.32), it can be shown that L_{ug} can be written as a linear composite of the X scores with the row vector of weights

$$\boldsymbol{b}'_g = \bar{\boldsymbol{X}}'_g S^{-1} \tag{3.34}$$

and the constant

$$c_g = -\frac{1}{2}\bar{\boldsymbol{X}}'_g S^{-1}\bar{\boldsymbol{X}}_g + \log q_g \tag{3.35}$$

As a result, L_{ug} can be written as $L_{ug} = \boldsymbol{b}'_g\boldsymbol{X}_u + c_g$. The weights defined by (3.34) and the constant values defined by (3.35) are usually reported in the output when running discriminant analysis with statistical software.

With these classification functions in mind, the maximum probability rule can be seen as: assign u to the population whose sample yields the largest QCF score (when covariance matrices are unequal) or the largest LCF score (when covariance matrices are equal). In terms of distance, the rule can be viewed as: assign u to the population to which it is "closest", where closeness is measured in terms of distances from observation vectors to sample centroids.

Looking at the QCF (3.29), it is easy to see that that maximizing Q_{ug} is equivalent

to minimizing

$$d_{ug} = -2Q_{ug} = \log|S_g| + D^2_{ug} - 2\log q_g \tag{3.36}$$

Therefore, the maximum probability, or in this case the minimum distance, rule may be written as:

$$\text{Assign unit } u \text{ to population } g \text{ if } d_{ug} < d_{ug'} \text{ for } g' \neq g \tag{3.37}$$

where d_{ug} is defined as in (3.36). This rule is equivalent to the rule stated in (3.30).

If it is the case where the covariance matrices are equal, then minimizing (3.36) is equivalent to minimizing

$$d^*_{ug} = D^{*2}_{ug} - 2\log q_g \tag{3.38}$$

In this case, the minimum distance rule may be written as:

$$\text{Assign unit } u \text{ to population } g \text{ if } d^*_{ug} < d^*_{ug'} \text{ for } g' \neq g \tag{3.39}$$

where d^*_{ug} is defined as in (3.38). This rule is equivalent to the rule stated in (3.33). The SAS DISCRIM program calls the expressions in (3.36) and (3.38) the *generalized squared distance function.*

To summarize, when in the realm of multivariate normal populations, the general statistic used in classification rules is that which involves the estimated posterior probabilities $\hat{P}(g|\boldsymbol{X})$ seen in (3.21), that which involves the quadratic classification function (QCF) (3.29), or that which involves the generalized distance seen in (3.36). All three of these statistics are equivalent and yield the same classification results. In the special case where the population matrices are equal, $\hat{P}(g|\boldsymbol{X})$ as seen in (3.27), the linear classification function (LCF) seen in (3.32), or the generalized distance seen in (3.38) are used. These three statistics are also equivalent, and they yield the same

classification results. The six major classification statistics are summarized below in Table 3.1 (Huberty, 1994).

Table 3.1: Classification Statistics

Quadratic

$$\hat{P}(g|\boldsymbol{X}_u) = \frac{q_g|S_g|^{-1/2}\exp\left(-\frac{1}{2}D^2_{ug}\right)}{\sum_{g'=1}^{k} q_{g'}|S_{g'}|^{-1/2}\exp\left(-\frac{1}{2}D^2_{ug'}\right)} \quad (3.21)$$

$$Q_{ug} = \log q_g - \frac{1}{2}\log|S_g| - \frac{1}{2}D^2_{ug} \quad (3.29)$$

$$d_{ug} = \log|S_g| + D^2_{ug} - 2\log q_g \quad (3.36)$$

Linear

$$\hat{P}(g|\boldsymbol{X}_u) = \frac{q_g\exp\left(-\frac{1}{2}D^{*2}_{ug}\right)}{\sum_{g'=1}^{k} q_{g'}\exp\left(-\frac{1}{2}D^{*2}_{ug'}\right)} \quad (3.27)$$

$$L_{ug} = (\bar{\boldsymbol{X}}'_g S^{-1})\boldsymbol{X}_u - \frac{1}{2}\bar{\boldsymbol{X}}'_g S^{-1}\bar{\boldsymbol{X}}_g + \log q_g \quad (3.32)$$

$$d^*_{ug} = D^{*2}_{ug} - 2\log q_g \quad (3.38)$$

Note: $D^2_{ug} = (\boldsymbol{X}_u - \bar{\boldsymbol{X}}_g)' S_g^{-1}(\boldsymbol{X}_u - \bar{\boldsymbol{X}}_g)$ (3.18)
$D^{*2}_{ug} = (\boldsymbol{X}_u - \bar{\boldsymbol{X}}_g)' S^{-1}(\boldsymbol{X}_u - \bar{\boldsymbol{X}}_g)$ (3.25)

4 Evaluating Classification Results

After performing predictive discriminant analysis (PDA), the next goal is to evaluate the accuracy of the results. Evaluating the accuracy of classification results involves estimating the true hit rates. There are three population hit rates, and they correspond to the following three questions: (1) How accurately can a classification rule based on population information be expected to perform? (2) How accurately can a classification rule based on the given sample be expected to classify units from future samples? (3) How accurately can a rule based on any sample of a fixed size be expected to classify units in future samples?

The first true hit rate is the *optimal hit rate*, denoted by $P^{(o)}$. This hit rate is obtained when a classification rule based on known parameters is applied to the population. The known parameters here are the k subpopulation mean vectors and common covariance matrix. The second hit rate is the *actual hit rate*, denoted by $P^{(a)}$. This hit rate is obtained by applying a classification rule based on one sample, often called the training sample, to future samples. The third true hit rate is the *expected actual hit rate*, denoted by $P^{(e)}$. This hit rate is the expected proportion of correct classifications over all possible samples. It is important to note that $P^{(e)} = E(P^{(a)})$.

4.1 Formula Estimators

In the special case where $k = 2$ and there is assumed multivariate normality with known population parameters (mean vectors and a common covariance matrix), it is possible to directly estimate $P^{(o)}$ and $P^{(a)}$ using formulas. The two population optimal hit rates are:

$$P_1^{(o)} = 1 - \phi\left(\frac{\Gamma - \Delta^2/2}{\Delta}\right), \; P_2^{(o)} = 1 - \phi\left(\frac{-\Gamma - \Delta^2/2}{\Delta}\right)$$

where ϕ is the standard normal distribution function ($\phi(x) = P(Z \leq x)$), $\Gamma = \log(\pi_2/\pi_1)$, π_g is the prior probability of membership in population g, and Δ^2 is the population Mahalanobis distance index. If $\pi_1 = \pi_2 = 1/2$, then $\Gamma = 0$ and $P^{(o)} = P_1^{(o)} = P_2^{(o)} = 1 - \phi(-\Delta/2) = \phi(\Delta/2)$. Sample values of Γ and Δ are used to calculate the estimate. The estimate of Γ is calculated by using research estimates of π_g, where $q_g = \hat{\pi}_g$. A nearly unbiased estimate of Δ^2 is $\tilde{D}^2$, where $\tilde{D}^2$ is defined as:

$$\tilde{D}^2 = \frac{N-p-3}{N-2} D^2 - \frac{pN}{n_1 n_2} \tag{4.1}$$

where p is the number of predictors and D^2 is the sample version of Δ^2. Therefore, the estimates of $P_1^{(o)}$ and $P_2^{(o)}$ are:

$$\hat{P}_1^{(o)} = 1 - \phi\left(\frac{K - \tilde{D}^2/2}{\tilde{D}}\right), \hat{P}_2^{(o)} = 1 - \phi\left(\frac{-K - \tilde{D}^2/2}{\tilde{D}}\right) \tag{4.2}$$

where $K = \hat{\Gamma} = \log(q_2/q_1)$. An estimator of the total population optimal hit rate is then:

$$\hat{P}^{(o)} = q_1 \hat{P}_1^{(o)} + q_2 \hat{P}_2^{(o)} \tag{4.3}$$

If $q_1 = q_2 = 1/2$, then $\hat{P}^{(o)} = \hat{P}_1^{(o)} = \hat{P}_2^{(o)} = 1 - \phi(-\tilde{D}/2) = \phi(\tilde{D}/2)$. No formula has been created for estimating the population optimal hit rate where $k > 2$ (Huberty, 1994).

A formula has also been developed to estimate the actual hit rate, $P^{(a)}$. This formula is also restricted to the two-group multivariate normal case. McLachlan (1975) developed a formula that provides estimates of $P_1^{(a)}$ and $P_2^{(a)}$. The McLachlan estimator for the gth ($g = 1, 2$) subpopulation actual hit rate is $\hat{P}_g^{(a)} = 1 - Q_g$, where

$$Q_g = \phi(-D/2) + f(-D/2)\Big\{\frac{p-1}{Dn_g} + \frac{D}{32m}[4(4p-1) - D^2] + \frac{(p-1)(p-2)}{4Dn_g^2}$$

$$+\frac{p-1}{64mn_g}[-D^3 + 8D(2p+1) + 16/D] + \frac{D}{12288m^2}[3D^6 - 4D^4(24p+7)$$

$$+16D^2(48p^2 - 48p - 53) + 192(-8p+15)]\Big\}$$

where f is the standard normal density function and $m = n_1 + n_2 - 2$. Several analyses have shown that for practical use, the last term in the multiplier of $f(-D/2)$ may be ignored. Therefore the shrinkage formula for $\hat{P}_g^{(a)}$ is:

$$\hat{P}_g^{(a)} = 1 - \phi(-D/2) + f(-D/2)\Big\{\frac{p-1}{Dn_g} + \frac{D}{32m}[4(4p-1) - D^2] + \frac{(p-1)(p-2)}{4Dn_g^2}$$

$$+\frac{p-1}{64mn_g}[-D^3 + 8D(2p+1) + 16/D]\Big\}$$

The estimate of the actual total population hit rate $P^{(a)}$ is then $\hat{P}^{(a)} = q_1\hat{P}_1^{(a)} + q_2\hat{P}_2^{(a)}$. Again, no formula has been created for $k > 2$.

4.2 Internal Analysis

When using a computer program to create classification results, the usual procedure is for the user to specify either a linear or a quadratic rule and the priors to be used. Then, the data are used to determine the k vectors of weights, $\boldsymbol{b}_g$, and the k constants, c_g. Next, these same data are used to determine the k QCF/LCF values for each unit, and then these data are used to reclassify the N units into the k populations. Following this process, the samples are classified based on parameters that are estimated from the samples themselves. This type of analysis is referred to as an *internal classification analysis*.

Results of internal analysis are presented as seen in Table 4.1, and Table 4.2 shows an example of a classification table.

Table 4.1: Classification Table for $k = 3$

		Predicted Group			
		1	2	3	Total
Actual Group	1	n_{11}	n_{12}	n_{13}	n_1
	2	n_{21}	n_{22}	n_{23}	n_2
	3	n_{31}	n_{32}	n_{33}	n_3
Total		$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	N

Table 4.2: Example Classification Table with $k = 3$

		Predicted Group			
		1	2	3	Total
Actual Group	1	42	0	3	45
	2	4	67	8	79
	3	0	2	60	62
Total		46	69	71	$N = 186$

In a classification table, $n_{gg'}$, located in cell (g, g'), is the number of units in group g that are classified as belonging to group g'. A *hit* is when a unit from group g is classified as belonging to group g. The *hit rate* for group g is defined as n_{gg}/n_g, and the total group hit rate is $\Sigma n_{gg}/N$. Looking at Table 4.2, the hit rate for group 1 is $42/45 = 93.3\%$, the hit rate for group 2 is $67/79 = 84.8\%$, the hit rate for group 3 is $60/62 = 96.8\%$, and the total group hit rate is $(42 + 67 + 60)/186 = 90.9\%$. These observed hit rates are called *apparent* or *resubstitution hit rates*. The apparent hit rate is used as a general hit rate estimate; it does not specifically estimate $P^{(o)}$,

$P^{(a)}$, or $P^{(e)}$ (Huberty, 1994). In SAS, internal results are labeled as "resubstitution" results.

4.3 External Analysis

In the internal analysis method described in the previous section, the units that are classified are the same as the units used to create the classification rules. Conversely, with *external classification analysis*, the classification rule is created from one set of units and then the rule is used to classify another set of units. This type of analysis follows the idea of *cross-validation*. Two types of external analyses are discussed, both of which involve splitting the sample.

4.3.1 Holdout Method

The first type of external analysis, the *holdout method*, involves a single splitting of the sample into two subsamples: (1) a *training* or design sample, and (2) a *test* or holdout sample. A classification rule is created using the training sample data, and then applied to the test sample data. A hit rate estimate is the proportion of the test sample units that are correctly classified. A holdout analysis may be performed using the SAS DISCRIM procedure. With this procedure, the user must externally generate the units to make up the training and test samples.

A few problems with the holdout method have been identified. First, it is recommended that the method is only used with large samples. Secondly, the classification rule is based on a portion of the sample, while ideally a classification rule should be based on the entire population. Thirdly, there are sometimes problems with determining the size of the test sample. If the test sample is too large, a good evaluation of the classification rule can be obtained, but the rule itself is most likely poor. On the other hand, if the test sample is too small, the classification rule will be better, but

it is harder to evaluate the accuracy of the rule. Lastly, the method is uneconomical with data. More data than is necessary to create a good classification rule is needed so that hit rate estimates can be calculated. Also, it is up to the researcher to decide what proportion of the total sample should be used as the test sample, and there is no general rule-of-thumb for determining this proportion (Huberty, 1994). For the $k = 2$ case, asymptotic theory developed by Schaafsma and van Vark (1979) suggests that the ratio of the test sample size to the training sample size is a function of p, the number of predictors, and is $[1 + (p-1)^{1/2}]^{-1}$. This led Schaafsma and van Vark to suggest a test-to-training ratio of 3/10 if there are at least 10 predictors. From this, it is generally accepted that using a test sample of 25-30% of the total sample is reasonable when $k = 2$. The estimated hit rate generated by the holdout method is not an appropriate estimator of $P^{(o)}$, $P^{(a)}$, or $P^{(e)}$. A holdout hit rate is only a good estimator of $P^{(a)}$ when the classification rule is considered to be based on a sample the size of the training sample, not the total original sample.

4.3.2 Leave-One-Out Method

The second method of external analysis is the leave-one-out (L-O-O) method proposed by Lachenbruch (1967). This method involves two steps. First, one unit is removed from the sample, and quadratic/linear classification functions are determined on the remaining $N - 1$ units. Next, these QCF/LCFs are used to classify the unit that had been removed into one of the k groups. This process is repeated N times, once for each of the N units. The hit rate estimates are the proportions of removed units that are correctly classified. In terms of the holdout method, it may be said that for each of the N steps in the process, there is a training sample size of $N - 1$ and a test sample of size 1. Hit rates based on the L-O-O method may be obtained using the SAS DISCRIM program. In SAS, L-O-O results are labeled as "cross validation"

results, and may be obtained by specifying CROSSVALIDATE as an option. As with hit rate estimates found using the holdout method, the L-O-O estimate is not an appropriate estimator of $P^{(o)}$, $P^{(a)}$, or $P^{(e)}$. The L-O-O estimate was originally designed to estimate $P^{(a)}$, but the estimated hit rate that this method provides is an estimate based on a sample size of $N-1$ rather than N. Despite this, unless N is small, the L-O-O estimator is a reasonable estimate of $P^{(a)}$.

Evidence has been presented to suggest a possible drawback with the leave-one-out method. In two separate simulation studies, Glick (1978) and Hora and Wilcox (1982) implied that the L-O-O method may provide hit rate estimates that have high variability over repeated sampling. This may be because of the reuse of the sample data since the N QCF/LCFs are derived from nearly identical samples. Despite this drawback, L-O-O estimates are easy to obtain, and the method is fairly robust to distribution violations, making them a quick and easy option for calculating hit rate estimates.

4.4 Maximum-Posterior-Probability Method

Most researchers use the holdout or the leave-one-out method when using cross-validation as a method for estimating hit rates. Another method was proposed by K. Fukunaga and D. L. Kessell (1971), and discussed by Glick (1978) and Hora and Wilcox (1982). This method is known as the maximum-posterior-probability (M-P-P) method. The M-P-P estimator for $P_g^{(a)}$ is a "mean" of the estimated posterior probabilities for units from all groups assigned to population g by the classification rule used. This estimator is called a "mean", because the sum of the estimated

posteriors is divided by Nq_g. The M-P-P estimator for $P_g^{(a)}$ is defined as:

$$\hat{P}_g^{(a)} = \frac{1}{Nq_g} \sum_{g'=1}^{k} \left[\sum_{u=1}^{n_{g'}} (\text{posterior probability for all } \boldsymbol{X}_u \text{ in group } g' \text{ assigned to group } g) \right] \tag{4.4}$$

The overall hit rate, $P^{(a)}$, can be estimated as follows:

$$\hat{P}^{(a)} = \sum_{g=1}^{k} q_g \hat{P}_g^{(a)} = \frac{1}{N} \sum_{u=1}^{N} \max[\hat{P}(1|\boldsymbol{X}_u), \hat{P}(2|\boldsymbol{X}_u), ..., \hat{P}(g|\boldsymbol{X}_u), ..., \hat{P}(k|\boldsymbol{X}_u)] \tag{4.5}$$

In this way, the estimate of $P^{(a)}$ is calculated from the mean of the maximum estimated posterior probabilities for each unit. The estimated posterior probabilities, $\hat{P}(g|\boldsymbol{X}_u)$, may be determined via either an internal analysis or an external analysis. When using internal estimated posterior probabilities, the estimator as denoted by M-P-P/I. When using external L-O-O estimated posterior probabilities, the estimator as denoted by M-P-P/L-O-O. In SAS, M-P-P estimates can be obtained in DISCRIM by specifying the POSTERR option. It is important to note that M-P-P estimates for separate groups, both stratified and unstratified, may be negative. This sometimes occurs because of discrepancies between group prior probabilities (q_g) and sample group proportions (n_g/N).

Evidence has been provided that show that M-P-P estimators have fairly good accuracy and reasonably good precision, meaning that they have low bias and low sampling variability. Hora and Wilcox (1982) conclude that if the multivariate normality assumption is valid, the M-P-P/L-O-O method is preferred to the L-O-O method. If the multivariate normality assumption is not valid, they concluded that the usual L-O-O method is better. They also conclude that the M-P-P/L-O-O estimator has greater accuracy than the M-P-P/I estimator. Monte Carlo results by Glick (1978) show that the M-P-P/I estimator has relatively low bias and lower sampling

variability than the L-O-O estimator for univariate prediction with two populations.

McLachlan (1992) and the SAS Institute (2011) call the M-P-P/L-O-O estimator a "smoothed" estimator. SAS (2011) calls the M-P-P/I and the M-P-P/L-O-O estimators *unstratified* (smoothed) estimators. SAS also defines a *stratified* estimator, which is an estimator that is stratified over the group from which the units emanate. The stratified estimator is defined as:

$$\hat{P}_g^{(a)}(\text{strat.}) = \frac{1}{q_g} \sum_{g'=1}^{k} \left[\frac{q_{g'}}{n_{g'}} \sum_{u=1}^{n_{g'}} (\text{post. prob. for all } \boldsymbol{X}_u \text{ in group } g' \text{ assigned to group } g) \right] \tag{4.6}$$

The choice between unstratified and stratified estimates depends on the confidence in the prior probability estimates. If the priors used are based on substantial knowledge of relative population sizes, the stratified estimates are preferred. If there is less confidence in the priors, then it is better to use the unstratified estimates. If the priors used are proportional to the group sizes within the sample, then the stratified and unstratified estimates will be the same.

5 Results

The data used for this analysis consist of 992 bat echolocation calls from two bat species: Indiana bat (*Myotis sodalis*) and little brown bat (*Myotis lucifugus*). For each call, the species of the bat from which the call originated is known. A summary of the data used is shown in Table 5.1. When choosing a bat species to compare

Table 5.1: Summary of the data

Bat Species	Number of Calls
Indiana Bat (*Myotis sodalis*)	408
Little Brown Bat (*Myotis lucifugus*)	584
Total	992

to the Indiana bat, the little brown bat was chosen because in past analyses of bat echolocation data, it was the bat species that was most often confused with the Indiana bat.

For the analyses that follow, three different statistical software packages are used: R, SAS, and SPSS. The choice to use multiple software packages was made for two reasons. First, by using different packages, results for the same type of analysis can be compared across packages. Secondly, some packages allow for types of analyses that others do not, therefore multiple software packages are used in an effort to perform a broad spectrum of analyses. For R, the functions qda() and lda() from the MASS package are used for quadratic and linear discriminant analysis, respectively. In SAS,

discriminant analysis is performed using proc DISCRIM. Lastly, discriminant analyis is performed in SPSS using DISCRIMINANT under the "classify" menu.

Linear and quadratic analyses are performed in both R and SAS, and a linear analysis is performed in SPSS. For both SAS and SPSS, all 76 response variables describing each call are used. In R, discriminant analysis using all 76 response variables is not possible because of collinearity. To alleviate this problem, a principle component analysis is performed using the princomp() function, and the first 67 principle components are used as the response variables in the discriminant analysis. The first 67 principle components are used because 67 is the maximum number of principle components for which the discriminant analysis will run.

Three main collections of analyses are performed. In the first batch of discriminant analyses, equal population prior probabilities are used. This means that if a bat call is randomly selected from the universe of bat calls, the probability that it belongs to an Indiana bat is equal to the probability that it belongs to a little brown bat. For the second batch of discriminant analyses, current population estimates of Indiana bats and little brown bats in the northeastern United States are used to construct population prior probabilities that reflect the distribution of these two bat populations in the wild. In the last set of analyses, linear discriminant analyses are performed along with stepwise variable selection in SPSS.

5.1 Discriminant Analysis using Equal Population Priors

5.1.1 Internal Analysis

The first type of analysis performed is an internal analysis. Linear and quadratic internal analyses are performed in both R and SAS, and a linear analysis is performed in SPSS. The classification tables resulting from these analyses can be seen in Tables

5.2, 5.3, and 5.4.

Looking at the three linear discriminant analyses, it is easy to see that both SAS and R provide the same classification results. The shared results from SAS and R have higher correct classification rates for both species (90.92% for the Indiana bat and 87.50% for the little brown bat) as well as a higher overall correct classification rate (89.52%) than the classification results from SPSS. Examining the two quadratic discriminant analyses, the SAS analysis has a better overall correct classification rate of 89.0% as well as a higher correct classification rate for the little brown bat of 94.86%. On the other hand, the analysis performed with R has a higher correct classification rate for the Indiana bat, and at 97.06%, it is very high.

Table 5.2: Classification Table from SAS Internal Discriminant Analysis

Quadratic				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	554 (94.86%)	30	584
	Indiana Bat	79	329 (80.64%)	408
	Total	633	359	992 (89.01%)
Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	531 (90.92%)	53	584
	Indiana Bat	51	357 (87.50%)	408
	Total	582	410	992 (89.52%)

Note: numbers in parentheses indicate correct classification rates

Table 5.3: Classification Table from SPSS Internal Discriminant Analysis

Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	523 (89.60%)	61	584
	Indiana Bat	64	344 (84.30%)	408
	Total	587	405	992 (87.40%)

Note: numbers in parentheses indicate correct classification rates

Table 5.4: Classification Table from R Internal Discriminant Analysis

Quadratic				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	444 (76.03%)	140	584
	Indiana Bat	12	396 (97.06%)	408
	Total	456	536	992 (89.01%)
Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	531 (90.92%)	53	584
	Indiana Bat	51	357 (87.50%)	408
	Total	582	410	992 (89.52%)

Note: numbers in parentheses indicate correct classification rates

5.1.2 Leave-One-Out Cross-Validation

After performing internal discriminant analyses, the next step is to evaluate the results. The first type of evaluation is the leave-one-out (L-O-O) cross-validation method. Linear and quadratic internal analyses are performed in both R and SAS, and a linear analysis is performed in SPSS. The L-O-O method can be performed in SAS by specifying the CROSSVALIDATE option, in R by specifying CV=TRUE inside the qda() and lda() functions, and in SPSS by checking off the box titled "leave-one-out classification" in the "classify" dialog box. The classification tables resulting from these analyses can be seen in Tables 5.5, 5.6, and 5.7.

Table 5.5: Classification Table from SAS L-O-O Cross-Validation

Quadratic				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	528 (90.41%)	56	584
	Indiana Bat	92	316 (77.45%)	408
	Total	620	372	992 (85.05%)
Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	523 (89.55%)	61	584
	Indiana Bat	65	343 (84.07%)	408
	Total	588	404	992 (87.30%)

Note: numbers in parentheses indicate correct classification rates

Table 5.6: Classification Table from SPSS L-O-O Cross-Validation

Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	514 (88.00%)	60	584
	Indiana Bat	73	335 (82.1%)	408
	Total	587	405	992 (85.60%)

Note: numbers in parentheses indicate correct classification rates

Table 5.7: Classification Table from R L-O-O Cross-Validation

Quadratic				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	416 (71.60%)	165	581
	Indiana Bat	27	378 (93.33%)	405
	Total	443	543	992 (80.53%)
Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	520 (89.04%)	64	584
	Indiana Bat	63	345 (84.56%)	408
	Total	583	409	992 (87.20%)

Note: numbers in parentheses indicate correct classification rates

Looking at the three linear discriminant analyses, it can be seen that the three software packages provide slightly different classification results. Of the three, SAS provides the highest overall correct classification rate at 87.30% as well as the highest correct classification rate for the little brown bat at 89.55%. The analysis performed by R gives the highest correct classification rate for the Indiana bat, 84.56%. Comparing the two quadratic discriminant analyses, SAS again provides the highest overall correct classification rate and the highest correct classification rate for the little brown bat at 85.08% and 90.41%, respectively. As with the linear analyses, R gives the higher correct classification rate for the Indiana bat (93.33%).

5.1.3 Holdout Cross-Validation

After using the leave-one-out cross-validation method, the next step is to perform a holdout analysis where larger portions of the data are used as the test sample. For this analysis, three different test sample sizes are used: 20%, 25%, and 33.33% of the total sample. The results from each of the three holdout analyses are arrived upon through a Monte Carlo simulation that consists of several steps. First, the sample is partitioned into the correct number of test samples. For the first analysis, where the test sample is 20% of the total sample, the sample was randomly divided into five mutually exclusive test samples. For the second analysis, where the test sample is 25% of the total sample, the sample is randomly divided into four mutually exclusive test samples. Lastly, for the third analysis where the test sample is 33.33% of the total sample, the sample is randomly divided into three mutually exclusive test samples. In the second step, a discriminant analysis is performed for each of the test samples. In each case, one test sample is used, and the training sample consists of the remainder of the total sample. For the first analysis, five discriminant analyses are performed, four for the second analysis, and three for the third analysis. To conclude this step,

the classification tables and correct classification rates for each of the discriminant analyses (five for the first analysis, four for the second analysis, and three for the third analysis) are averaged. The result is one classification table and one set of correct classification rates per analysis. The entire process is then repeated $N = 1000$ times for each analysis. The results, seen in Tables 5.8, 5.9, and 5.10, are the averages from the $N = 1000$ repetitions.

Looking at the results of the three holdout analyses, there are several patterns that can be seen. For the linear discriminant analyses, the highest overall and group correct classification rates (86.86% overall, 89.20% for little brown bats, and 83.51% for Indiana bats) and the lowest standard error can be found when the size of the test sample is the smallest. As the test samples get larger, the classification rates get lower and the standard errors get higher. The quadratic discriminant analyses show the opposite behavior for the overall correct classification rate and the correct

Table 5.8: Classification Table from Holdout Cross-Validation, Test Sample = 20%

Quadratic			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	86.0212 (73.65%)	30.7788
	Indiana Bat	7.2112	74.3888 (91.16%)
		Overall Rate = 80.85%, SE = 0.000219	
Linear			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	104.1804 (89.20%)	12.6196
	Indiana Bat	13.4530	68.1470 (83.51%)
		Overall Rate = 86.86%, SE = 0.000144	

Note: numbers in parentheses indicate correct classification rates, N=1000

Table 5.9: Classification Table from Holdout Cross-Validation, Test Sample = 25%

Quadratic			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	108.6800 (74.44%)	37.3200
	Indiana Bat	9.7255	92.2745 (90.47%)
		Overall Rate = 81.03%, SE = 0.000277	
Linear			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	130.1125 (89.12%)	15.8875
	Indiana Bat	16.8765	85.1235 (83.45%)
		Overall Rate = 86.79%, SE = 0.000163	

Note: numbers in parentheses indicate correct classification rates, N=1000

Table 5.10: Classification Table from Holdout Cross-Validation, Test Sample = 33%

Quadratic			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	148.02967 (76.04%)	46.6370
	Indiana Bat	14.60333	121.3967 (89.26%)
		Overall Rate = 81.48%, SE = 0.000367	
Linear			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	173.114000 (88.93%)	21.55267
	Indiana Bat	22.7067	113.29233 (83.30%)
		Overall Rate = 86.61%, SE = 0.000195	

Note: numbers in parentheses indicate correct classification rates, N=1000

classification rate for little brown bats. The highest rates (81.48% overall and 76.04% for little brown bats) are found when the test sample is the largest, and these rates get smaller as the test sample size decreases. Conversely, the highest correct classification rate for Indiana bats (91.16%) and the lowest standard error are found when the test sample is the smallest. The rate for Indiana bats decreases and the standard error increases as the size of the test sample increases.

5.1.4 Maximum-Posterior-Probability Estimates

The last type of evaluation to perform is to use the maximum-posterior probability (M-P-P) method to calculate classification error rates. This is done in SAS by specifying the POSTERR option. For internal estimates, M-P-P/I estimates are calculated, and for external estimates, M-P-P/L-O-O estimates are calculated. For both external and internal estimates, both stratified and unstratified estimates are calculated. The posterior probability error rate estimates can be seen in Tables 5.11 and 5.12. It is important to note that the rates seen in these tables are error rates; the correct classification rates are the complements of these error rates. In the case where the error rate is negative, the correct classification rate is said to be equal to 1.000 (100%).

Table 5.11: M-P-P/I Error Rate Estimates

Quadratic			
	Little Brown Bat	Indiana Bat	Total
Stratified	-0.1298	0.1516	0.0109
Unstratified	-0.2652	0.2846	0.0097
Linear			
	Little Brown Bat	Indiana Bat	Total
Stratified	0.0648	0.1367	0.1008
Unstratified	-0.0664	0.2665	0.1001

Table 5.12: M-P-P/L-O-O Error Rate Estimates

Quadratic			
	Little Brown Bat	Indiana Bat	Total
Stratified	-0.1163	0.1386	0.0111
Unstratified	-0.2376	0.2584	0.0104
Linear			
	Little Brown Bat	Indiana Bat	Total
Stratified	0.0518	0.1565	0.1041
Unstratified	-0.0708	0.2781	0.1036

When comparing the stratified estimated error rates to the unstratified error rates in these tables, it is important to note that for these analyses, equal population prior probabilities were used. Since these priors did not come from knowledge of the relative population sizes, the unstratified estimates are preferred. First look at the M-P-P/I estimated error rates. All error rates for the little brown bat, except for the linear stratified estimate, are negative. For the Indiana bat, the linear error rate estimates are smaller than their quadratic counterparts. Lastly, for the total estimated error rates, the quadratic estimates are smaller than the corresponding linear estimates. Looking at the M-P-P/L-O-O estimates, again it can be seen that all error rates for the little brown bat, except for the linear stratified estimate, are negative. For the Indiana bat, the quadratic estimates are smaller than the corresponding linear error rate estimates. For the total estimated error rate, the quadratic analysis provides much smaller error rate estimates.

5.2 Discriminant Analysis using Unequal Population Priors

The second collection of analyses is a set of discriminant analyses performed with unequal population prior probabilities. The prior probabilities are calculated from current population estimates of the Indiana bat and the little brown bat in the north-

eastern United States. The Indiana bat population is currently estimated to be 54,142 (U.S. Fish and Wildlife Service, b), and the little brown bat population is currently estimated to be 650,000 (Frick et al., 2010). Using these population estimates, the population priors are found to be 0.9231 for little brown bats and 0.07690 for Indiana bats.

5.2.1 Internal Analysis

The first analysis performed is an internal analysis. Both quadratic and linear analyses are performed in SAS and R, and the results can be seen in Tables 5.13 and 5.14.

Table 5.13: Classification Table from SAS Internal Discriminant Analysis

Quadratic				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	559 (95.72%)	25	584
	Indiana Bat	90	318 (77.94%)	408
	Total	649	343	992 (88.41%)
Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	577 (98.80%)	7	584
	Indiana Bat	175	233 (57.11%)	408
	Total	752	240	992 (81.65%)

Note: numbers in parentheses indicate correct classification rates

Comparing the two quadratic analyses, SAS provides the higher correct classification rates overall and for little brown bats (88.41% and 95.72% respectively), while R provides the higher correct classification rate for Indiana bats (95.83%). Looking

Table 5.14: Classification Table from R Internal Discriminant Analysis

Quadratic				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	473 (80.99%)	111	584
	Indiana Bat	17	391 (95.83%)	408
	Total	490	502	992 (87.10%)
Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	577 (98.80%)	7	584
	Indiana Bat	178	230 (56.37%)	408
	Total	755	237	992 (81.35%)

Note: numbers in parentheses indicate correct classification rates

at the two linear analyses, SAS provides higher correct classification rates in all three categories (81.65% overall, 98.80% for little brown bats, and 57.11% for Indiana bats). It is interesting to note that in the linear analyses, the correct classification rates for Indiana bats are very low.

5.2.2 Leave-One-Out Cross-Validation

The second analysis performed is discriminant analysis with leave-one-out (L-O-O) cross-validation. Both quadratic and linear discriminant analyses are performed in SAS and R, and the results can be seen in Tables 5.15 and 5.16. Looking at the two quadratic analyses, SAS provides the higher correct classification rates overall and for the little brown bat (84.78% and 91.44% respectively), while R provides the higher correct classification rate for Indiana bats (91%). When comparing these two

analyses, it is important to note that they provide very different classifications. SAS does a very good job classifying little brown bats, but an adequate job classifying Indiana bats. Conversely, R does a very good job classifying Indiana bats, but only a decent job classifying little brown bats. Now looking at the two linear analyses, it can immediately be seen that the two classifications are very similar. However, SAS provides slightly higher rates in all three categories (80.04% overall, 97.77% for little brown bats, and 54.66% for Indiana bats). Both SAS and R do an extremely good job classifying little brown bats, but a poor job classifying Indiana bats.

Table 5.15: Classification Table from SAS L-O-O Cross-Validation

Quadratic				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	534 (91.44%)	50	584
	Indiana Bat	101	307 (75.25%)	408
	Total	635	357	992 (84.78%)
Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	571 (97.77%)	13	584
	Indiana Bat	185	223 (54.66%)	408
	Total	756	236	992 (80.04%)

Note: numbers in parentheses indicate correct classification rates

Table 5.16: Classification Table from R L-O-O Cross-Validation

Quadratic				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	434 (74.70%)	147	581
	Indiana Bat	33	372 (91.60%)	405
	Total	467	519	992 (81.74%)
Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	570 (97.60%)	14	584
	Indiana Bat	186	222 (54.41%)	408
	Total	756	236	992 (79.84%)

Note: numbers in parentheses indicate correct classification rates

5.2.3 Holdout Cross-Validation

The third type of analysis performed is discriminant analysis with holdout cross-validation. The same process that is described in the previous section is used here, and the results can be seen in Tables 5.17, 5.18, and 5.19. Looking at the quadratic analyses, the highest classification rates are all found when the test sample size is the largest. These rates are 82.42% overall, 78.40% for little brown bats, and 92.24% for Indiana bats. However, the smallest standard error is found when the test sample size is the smallest. All three quadratic analyses provide similar classifications; they all do a very good job classifying Indiana bats but merely an adequate job classifying little brown bats.

Comparing the linear analyses, the highest correct classification rate for little brown bats (97.47%) and the lowest standard error occur when the test sample size

Table 5.17: Classification Table from Holdout Cross-Validation, Test Sample = 20%

Quadratic			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	89.578 (76.69%)	27.222
	Indiana Bat	8.008	73.592 (90.19%)
		Overall Rate = 82.24%, SE = 0.000206	
Linear			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	113.8496 (97.47%)	2.9504
	Indiana Bat	36.6512	44.9488 (55.08%)
		Overall Rate = 80.04%, SE = 0.000133	

Note: numbers in parentheses indicate correct classification rates, N=1000

Table 5.18: Classification Table from Holdout Cross-Validation, Test Sample = 25%

Quadratic			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	112.8240 (77.28%)	33.1760
	Indiana Bat	10.7505	91.2495 (88.59%)
		Overall Rate = 82.29%, SE = 0.000264	
Linear			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	142.2252 (97.41%)	3.77475
	Indiana Bat	45.5750	56.4250 (55.32%)
		Overall Rate = 80.10%, SE = 0.000144	

Note: numbers in parentheses indicate correct classification rates, N=1000

Table 5.19: Classification Table from Holdout Cross-Validation, Test Sample = 33%

Quadratic			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	152.61867 (78.40%)	42.0480
	Indiana Bat	10.08667	92.24 (92.24%)
		Overall Rate = 82.42%, SE = 0.000371	
Linear			
		Predicted Group	
		Little Brown Bat	Indiana Bat
Actual Group	Little Brown Bat	189.45133 (97.32%)	5.215333
	Indiana Bat	60.10833	75.891667 (55.80%)
		Overall Rate = 80.24%, SE = 0.000169	

Note: numbers in parentheses indicate correct classification rates, N=1000

is the smallest. Conversely, the highest correct classification rates overall and for Indiana bats (80.24% and 55.80% respectively) occur when the test sample size is the largest. Despite the slight differences, all three linear analyses provide similar classifications, where little brown bats are classified very well and Indiana bats are classified very poorly.

5.2.4 Maximum-Posterior-Probability Estimates

The last type of analysis is the calculation of maximum-posterior-probability (M-P-P) error rates. Like before, both stratified and unstratified estimates are calculated for M-P-P/I and M-P-P/L-O-O error rates. The estimated error rates can be seen in Tables 5.20 and 5.21. This time, since the population prior probabilities are calculated using current bat population estimates, the stratified estimates are preferred. Looking at the M-P-P/I error rates, all of the error rates for little brown bats are positive,

and all but one of the rates for Indiana bats are negative. In all cases, the quadratic estimates are smaller than their linear counterparts. Comparing the M-P-P/L-O-O error rate estimates, all of the error rates for little brown bats are positive, and all but one of the rates for Indiana bats are negative. The quadratic total error rate estimates and the error rate estimates for Indiana bats are lower than the corresponding linear estimates. However, the linear error rate estimates for little brown bats are smaller than the quadratic error rate estimates.

Table 5.20: M-P-P/I Error Rate Estimates

Quadratic			
	Little Brown Bat	Indiana Bat	Total
Stratified	0.0279	-0.2745	0.0046
Unstratified	0.2965	-3.4592	0.0077
Linear			
	Little Brown Bat	Indiana Bat	Total
Stratified	0.0060	0.4180	0.0377
Unstratified	0.2307	-1.6922	0.0828

Table 5.21: M-P-P/L-O-O Error Rate Estimates

Quadratic			
	Little Brown Bat	Indiana Bat	Total
Stratified	0.0687	-0.7355	0.0069
Unstratified	0.3114	-3.6167	0.0094
Linear			
	Little Brown Bat	Indiana Bat	Total
Stratified	0.0159	0.3403	0.0408
Unstratified	0.2261	-1.6401	0.0826

5.3 Discriminant Analysis with Variable Selection

As a last form of analysis, two discriminant analyses are performed in SPSS using stepwise variable selection. An internal discriminant analysis and a discriminant anal-

ysis with leave-one-out (L-O-O) cross-validation are performed with equal population prior probabilities. For these analyses, SPSS selected the following 18 variables: TotalSlope, LedgeDuration, RelPwr2ndTo1st, SlopeAtFc, FreqLedge, DurOf5dB, StartSlope, FreqCtr, HiFtoFcDmp, AmpK@start, FreqMaxPwr, HiFtoKnSlope, Amp3rdQrtl, Amp3rdMean, AmpKurtosis, AmpEndLn60ExpC, FcMinusEndF, and Amp4thQrtl. The classification tables resulting from these two analyses can be seen in Tables 5.22 and 5.23.

Table 5.22: Classification Table from SPSS Internal Discriminant Analysis

Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	533 (91.30%)	51	584
	Indiana Bat	63	345 (84.6%)	408
	Total	596	396	992 (88.50%)

Note: numbers in parentheses indicate correct classification rates

Table 5.23: Classification Table from SPSS L-O-O Cross-Validation

Linear				
		Predicted Group		
		Little Brown Bat	Indiana Bat	Total
Actual Group	Little Brown Bat	531 (90.90%)	53	584
	Indiana Bat	66	342 (83.8%)	408
	Total	597	395	992 (88.00%)

Note: numbers in parentheses indicate correct classification rates

Comparing Table 5.22 to the corresponding linear internal analyses seen in Tables 5.2, 5.3, and 5.4, it is easy to see that by reducing the number of response variables,

SPSS is able to increase all three correct classification rates for its own analysis (Table 5.3), as well improve upon the correct classification rate for little brown bats provided by SAS and R. Comparing Table 5.23 to the corresponding linear L-O-O analyses seen in Tables 5.5, 5.6, and 5.7, it can be seen that through using stepwise variable selection, SPSS is able to improve all three correct classification rates for its own analysis (Table 5.6) and improve upon the overall correct classification rates and the correct classification rates for little brown bats provided by SAS and R.

6 Conclusions

The Indiana bat, *Myotis sodalis*, is an endangered species of bat found in the eastern half of the United States. Since the start of the White Nose Syndrome epidemic in 2006, the already declining population of Indiana bats has been decimated. In an effort to avoid the extinction of the species, the location and identification of Indiana bats in the wild has become a priority. One of the methods for locating and identifying bats is through the collection and analysis of echolocation calls. Acoustically, Indiana bats and little brown bats (*Myotis lucifugus*) are very similar. The goal of this study was therefore to determine if acoustic methods are a reliable way to discriminate between these two bat species.

The data used for this study consist of 992 echolocation calls from Indiana bats and little brown bats. For each of the calls, 76 parameters were measured by SonoBat. Quadratic and linear discriminant analyses were performed, each using a different statistical software and/or evaluation method. All of these analyses were performed using equal population prior probabilities and unequal population prior probabilities. Tables 6.1 and 6.2 compare the correct classification rates determined by each of these analyses.

The major finding of this study is that the majority of the correct classification rates fall in the range of 85%-90%. Given the large amount of calls and response variables involved in these analyses, rates this low indicate that the two bat species are extremely similar acoustically, and are therefore hard to discriminate from each other. This means that if acoustic methods were used to differentiate between Indiana and little brown bats in the wild, researchers would be dealing with error rates in the range of 10%-15%. Given the specific circumstances of locating and identifying Indiana bats, many scientists would agree that these error rates are too high.

Table 6.1: Comparison of Correct Classification Rates using Equal Prior Probabilities

Method	Overall	Indiana Bat	Little Brown Bat
SAS Internal quadratic	89.01%	80.64%	94.86%
SAS Internal linear	89.52%	87.50%	90.92%
SPSS Internal linear	87.40%	84.30%	89.60%
SPSS Internal linear*	88.50%	84.60%	91.30%
R Internal quadratic	89.01%	97.06%	76.03%
R Internal linear	89.52%	87.50%	90.92%
SAS L-O-O quadratic	85.05%	77.45%	90.41%
SAS L-O-O linear	87.30%	84.07%	89.55%
SPSS L-O-O linear	85.60%	82.10%	88.00%
SPSS L-O-O linear*	88.00%	83.80%	90.90%
R L-O-O quadratic	80.53%	93.33%	71.60%
R L-O-O linear	87.20%	84.56%	89.04%
Holdout (Test = 20%) quadratic	80.85%	91.16%	73.65%
Holdout (Test = 20%) linear	86.86%	83.51%	89.20%
Holdout (Test = 25%) quadratic	81.03%	90.47%	74.44%
Holdout (Test = 25%) linear	86.79%	83.45%	89.12%
Holdout (Test = 30%) quadratic	81.48%	89.26%	76.04%
Holdout (Test = 30%) linear	86.61%	83.30%	88.93%
M-P-P/I quadratic**	99.03%	71.54%	100%***
M-P-P/I linear**	89.99%	73.35%	100%***
M-P-P/L-O-O quadratic**	98.96%	74.16%	100%***
M-P-P/L-O-O linear**	89.64%	72.19%	100%***

* discriminant analysis using variable selection
** unstratified error rate estimate
*** error rate estimate is negative

Looking more closely at the correct classification rates seen in Tables 6.1 and 6.2, more observations can be made. First, the internal and L-O-O rates from the quadratic discriminant analyses (QDA) performed by R are particularly different from the rates calculated by SAS and SPSS; the rates for Indiana bats are much higher and the rates for little brown bats are much lower. It is also important to note that there is much more variation in the correct classification rates when unequal prior probabilities are used. When this is the case, analyses tend to do well for one bat species but not the other, meaning that the rate for one of the two bat species tends

Table 6.2: Comparison of Correct Classification Rates using Unequal Prior Probabilities

Method	Overall	Indiana Bat	Little Brown Bat
SAS Internal quadratic	88.41%	77.94%	95.72%
SAS Internal linear	81.65%	57.11%	98.80%
R Internal quadratic	87.10%	95.88%	80.99%
R Internal linear	81.35%	56.37%	98.80%
SAS L-O-O quadratic	84.78%	75.25%	91.44%
SAS L-O-O linear	80.04%	54.66%	97.77%
R L-O-O quadratic	81.74%	91.60%	74.70%
R L-O-O linear	79.84%	54.41%	97.60%
Holdout (Test = 20%) quadratic	82.24%	90.19%	76.69%
Holdout (Test = 20%) linear	80.04%	55.08%	97.47%
Holdout (Test = 25%) quadratic	82.29%	88.59%	77.28%
Holdout (Test = 25%) linear	80.10%	55.32%	97.41%
Holdout (Test = 30%) quadratic	82.42%	92.24%	78.40%
Holdout (Test = 30%) linear	80.24%	55.80%	97.32%
M-P-P/I quadratic*	99.54%	100%**	97.21%
M-P-P/I linear*	96.23%	58.20%	99.40%
M-P-P/L-O-O quadratic*	99.31%	100%**	93.13%
M-P-P/L-O-O linear*	95.92%	65.97%	98.41%

* stratified error rate estimate

** error rate estimate is negative

to be much higher than the rate for the other.

Comparing the results from the three different holdout analyses with equal prior probabilities, there are two major trends. Generally, as the size of the test sample increases, the QDA rates increase and the linear discriminant analysis (LDA) rates decrease. The rates from the holdout analyses with unequal prior probabilities do not change as much as the size of the test sample increases. However, there is a major difference between the QDA and LDA rates when unequal prior probabilities are used. With QDA, the rates for Indiana bats are very high and with LDA they are very low. Conversely, the LDA rates are very high for little brown bats while the QDA rates are low.

Appendix A Parameters Measured by SonoBat

Table A.1: Parameters measured by the SonoBat detector (SonoBat)

TimeInFile	Time of the call within the file (milliseconds).
PrecedingIntrvl	Time between the current call and the previous call (milliseconds).
CallsPerSec	Mean calls per second of the recording or section of recording displayed.
CallDuration	Duration of the call (milliseconds).
Fc	Characteristic frequency of the call. This is determined by finding the point in the final 40% of the call having the lowest slope or exhibiting the end of the main trend of the body of the call (kHx).
HiFreq	Highest apparent frequency of the call.
LowFreq	Lowest apparent frequency of the call.

Bndwdth	Total frequency spread of the call. This is calculated from the difference between the highest and lowest frequency.
FreqMaxPwr	The frequency of the maximum amplitude of the call.
PrcntMaxAmpDur	Percentage of the entire call duration at which the maximum amplitude occurs.
TimeFromMaxToFc	Time from the point at which the maximum amplitude occurs to the point in the call of the characteristic frequency.
FreqKnee	Frequency at which the initial slope of the call most abruptly transitions to the slope of the body of the call.
PrcntKneeDur	Percentage of the entire call duration at which the knee occurs, i.e., the point at which the initial slope of the call most abruptly transitions to the slope of the body of the call.
StartF	Frequency of the start of the call. This is typically the same point as the highest frequency, but it will be different if the call initially rises in frequency.
EndF	Frequency of the end of the call. This is typically the same point as the lowest frequency, but it will be different if the call ends with a rise in frequency.

DominantSlope	Slope of the longest sustained trend in slope of the call. This is determined by finding the segment of the call having the minimum residue for a linear regression of a segment of the call of 20% the duration of the call (kHx/msec).
SlopeAtFc	Instantaneous slope at the point of the characteristic frequency.
StartSlope	Slope at the start of the call. This is calculated from the first 5% of the call duration.
EndSlope	Slope at the end of the call. This is calculated from the final 5% of the call duration.
SteepestSlope	Steepest slope of the call. This is calculated from a linear regression of a segment of 10% the duration of the call.
LowestSlope	Lowest slope of the call. This is calculated from a linear regression of a segment of 10% the duration of the call.
TotalSlope	Total slope of the call. This is calculated from the difference in frequency and time from the point of highest frequency to the point of the characteristic frequency.
HiFtoKnSlope	Slope of the call calculated from the difference in frequency and time from the point of highest frequency to the point of the knee.
KneeToFcSlope	Slope of the call calculated from the difference in frequency and time from the point of the knee to the point of the characteristic frequency.

CummNmlzSlp	Average of the instantaneous slopes of the call.
HiFtoFcExpAmp	Amplitude parameter of an exponential fit of the call from the point of high frequency to the point of the characteristic frequency.
HiFtoFcDmp	Damping parameter of an exponential fit of the call from the point of high frequency to the point of the characteristic frequency.
KnToFcExpAmp	Amplitude parameter of an exponential fit of the call from the point of the knee to the point of the characteristic frequency.
KnToFcDmp	Damping parameter of an exponential fit of the call from the point of the knee to the point of the characteristic frequency.
HiFtoKnExpAmp	Amplitude parameter of an exponential fit of the call from the point of the high frequency to the point of the characteristic frequency.
HiFtoKnDmp	Damping parameter of an exponential fit of the call from the point of the high frequency to the point of the characteristic frequency.
FreqLedge	Frequency of the ledge, i.e., the most abrupt transition to the most extended flattest slope section of the body of the call preceding the characteristic frequency. This is also called the "ledge" of the call.

LedgeDuration	Duration of the ledge, i.e., the most extended flattest slope section of the body of the call preceding the characteristic frequency.
FreqCtr	Frequency at the center of the duration of the call.
FBak32dB	Frequency of the call 32 dB below the point of maximum amplitude of the call, and preceding the point of maximum amplitude of the call.
FFwd32dB	Frequency of the call 32 dB below the point of maximum amplitude of the call, and after the point of maximum amplitude of the call.
FBak20dB	Frequency of the call 20 dB below the point of maximum amplitude of the call, and preceding the point of maximum amplitude of the call.
FFwd20dB	Frequency of the call 20 dB below the point of maximum amplitude of the call, and after the point of maximum amplitude of the call.
FBak15dB	Frequency of the call 15 dB below the point of maximum amplitude of the call, and preceding the point of maximum amplitude of the call.
FFwd15dB	Frequency of the call 15 dB below the point of maximum amplitude of the call, and after the point of maximum amplitude of the call.

FBak5dB	Frequency of the call 5 dB below the point of maximum amplitude of the call, and preceding the point of maximum amplitude of the call.
FFwd5dB	Frequency of the call 5 dB below the point of maximum amplitude of the call, and after the point of maximum amplitude of the call.
Bndw32dB	The total bandwidth covered from the point of the call 32 dB below and before the point of maximum amplitude and the point of the call 32 dB below and after the point of maximum amplitude of the call.
Bndw20dB	The total bandwidth covered from the point of the call 20 dB below and before the point of maximum amplitude and the point of the call 32 dB below and after the point of maximum amplitude of the call.
Bndw15dB	The total bandwidth covered from the point of the call 15 dB below and before the point of maximum amplitude and the point of the call 32 dB below and after the point of maximum amplitude of the call.
Bndw5dB	The total bandwidth covered from the point of the call 5 dB below and before the point of maximum amplitude and the point of the call 32 dB below and after the point of maximum amplitude of the call.
DurOf32dB	The duration of the call from the point of the call 32 dB below and before the point of maximum amplitude and the point of the call 32 dB below and after the point of maximum amplitude of the call.

DurOf20dB	The duration of the call from the point of the call 20 dB below and before the point of maximum amplitude and the point of the call 32 dB below and after the point of maximum amplitude of the call.
DurOf15dB	The duration of the call from the point of the call 15 dB below and before the point of maximum amplitude and the point of the call 32 dB below and after the point of maximum amplitude of the call.
DurOf5dB	The duration of the call from the point of the call 5 dB below and before the point of maximum amplitude and the point of the call 32 dB below and after the point of maximum amplitude of the call.
Amp1stQrtl	Total amplitude of the first quartile of the call (relative units).
Amp2ndQrtl	Total amplitude of the second quartile of the call (relative units).
Amp3rdQrtl	Total amplitude of the third quartile of the call (relative units).
Amp4thQrtl	Total amplitude of the fourth quartile of the call (relative units).
Amp1stMean	Mean of the first quartile amplitude (relative units).
Amp2ndMean	Mean of the second quartile amplitude (relative units).

Amp3rdMean	Mean of the third quartile amplitude (relative units).
Amp4thMean	Mean of the fourth quartile amplitude (relative units).
LnExpA_StartAmp	Amplitude parameter of an exponential fit of the time-amplitude trend of the call from the start of the call to the point of maximum amplitude.
LnExpB_StartAmp	Damping parameter of an exponential fit of the time-amplitude trend of the call from the start of the call to the point of maximum amplitude.
AmpStartLn60ExpC	Time parameter of an exponential fit of the time-amplitude trend of the call from the start of the call to the point of maximum amplitude.
LnExpA_EndAmp	Amplitude parameter of an exponential fit of the time-amplitude trend of the call from the point of maximum amplitude to the end of the call.
LnExpB_EndAmp	Damping parameter of an exponential fit of the time-amplitude trend of the call from the point of maximum amplitude to the end of the call.
AmpEndLn60ExpC	Time parameter of an exponential fit of the time-amplitude trend of the call from the point of maximum amplitude to the end of the call.
AmpK@start	Slope of a logarithmic plot of the time-amplitude trend of the call from the start of the call to the point of maximum amplitude.

AmpK@end	Slope of a logarithmic plot of the time-amplitude trend of the call from the point of maximum amplitude to the end of the call.
AmpKurtosis	Kurtosis of the time-amplitude trend.
AmpSkew	Skew of the time-amplitude trend.
AmpVariance	Variance of the time-amplitude trend.
AmpMoment	Moment of the time-amplitude trend.
AmpGausR2	R-squared of a Gaussian fit of the time amplitude trend.
Quality	Quality rating of the call based on the total points of the sonogram above a threshold value. SonoBat uses this synthesized measure to assist in the call trending analysis of strong and weak call signals.
HiFminusStartF	High frequency minus start frequency.
FcMinusEndF	Characteristic frequency minus start frequency.
RelPwr2ndTo1st	Ratio of the strength of the harmonic that SonoBat trended to the strength of the next higher harmonic.
RelPwr3rdTo1st	Ratio of the strength of the harmonic that SonoBat trended to the strength of the second higher harmonic.

References

Eric R. Britzke, Kevin L. Murray, John S. Heywood, and Lynn W. Robbins. Acoustic identification. In A. Kurta and J. Kennedy, editors, *The Indiana Bat: Biology and Management of an Endangered Species*, pages 221–225, Austin, TX, 2002. Bat Conservation International.

Eric R. Britzke, Joseph E. Duchamp, Kevin L. Murray, Robert K. Swihart, and Lynn W. Robbins. Acoustic identification of bats in the eastern united states: A comparison of parametric and nonparametric methods. *Journal of Wildlife Management*, 75(3):660–667, 2011.

Jeffrey P. Cohn. White-nose symdrome threatens bats. *BioScience*, 58(11):1098, 2008.

M. Brock Fenton. Picking the right bat detector - time expansion versus zero-crossing. *Bat Research News*, 41(4):116, 2000.

Janet Foley, Deana Clifford, Kevin Castle, Paul Cryan, and Richard S. Ostfeld. Investigating and managing the rapid emergence of white-nose syndrome, a novel, fatal, infectious disease of hibernating bats. *Conservation Biology*, 25(2):223–231, 2011.

Winifred F. Frick, Jacob F. Pollock, Alan C. Hicks, Kate E. Langwig, D. Scott Reynolds, Gregory G. Turner, Calvin M. Butchkoski, and Thomas H. Kunz. An emerging disease causes regional population collapse of a common north american bat species. *Science*, 319(5992):679–682, 2010.

K. Fukunaga and D. L. Kessell. Estimation of classification error. *Computers, IEEE Transactions on*, C-20(12):1521–1527, 1971.

William L. Gannon and Richard E. Sherwin. Are acoustic detectors a 'Silver Bullet' for assessing habitat use by bats? In R. Mark Brigham, Elisabeth K. V. Kalko, Gareth Jones, Stuart Parsons, and Herman J. G. A. Limpens, editors, *Bat Echolocation Research: Tools, Techniques, and Analysis*, pages 38–45, Miami, FL, 2000. North American Symposium on Bat Research.

N. Glick. Additive estimators for probabilities of correct classification. *Pattern Recognition*, 10:211–227, 1978.

S. C. Hora and J. B. Wilcox. Estimation of error rates in several population discriminant analysis. *Journal of Marketing Research*, 19:57–61, 1982.

Carl J. Huberty. *Applied Discriminant Analysis*. John Wiley and Sons, Inc., New York, 1994.

Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, 6th edition, 2007.

Amy J. Kuenzi and Michael L. Morrison. Detection of bats by mist-nets and ultrasonic sensors. *Wildlife Society Bulletin*, 26(2):307–311, 1998.

P. A. Lachenbruch. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics*, 23: 639–645, 1967.

G. J. McLachlan. Confidence intervals for the conditional probability of misallocation in discriminant analysis. *Biometrics*, 31:161–167, 1975.

Geoffrey J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley and Sons, Inc., New York, 1992.

Michael J. O'Farrell and William L. Gannon. A comparison of acoustic versus capture techniques for the inventory of bats. *Journal of Mammalogy*, 80(1):24–30, 1999.

Stuart Parsons and Gareth Jones. Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artificial neural networks. *The Journal of Experimental Biology*, 203:2641–2656, 2000.

Lars Pettersson. The properties of sound and bat detectors. In Gareth Jones Stuart Parsons R. Mark Brigham, Elisabeth K. V. Kalko and Herman J. G. A. Limpens, editors, *Bat Echolocation Research: Tools, Techniques, and Analysis*, pages 9–13, Miami, FL, 2000. North American Symposium on Bat Research.

Damiano G. Preatoni, Mose Docari, Roberta Chirichella, Guido Tosi, Luc A. Wauters, and Adriano Martinoli. Identifying bats from time-expanded recordings of search calls: Comparing classification methods. *Journal of Wildlife Management*, 69(4): 1601–1614, 2005.

C. R. Rao. *Linear Statistical Inference and its Applications*. John Wiley and Sons, Inc., New York, 2nd edition, 1973.

Danilo Russo and Gareth Jones. Identification of twenty-two bat species (mammalia: Chiroptera) from italy by analysis of time-expanded recordings of echolocation calls. *Journal of Zoology*, 258(1):91–103, 2002.

Colin Campbell Sanborn. Bats of the united states. *Public Health Reports*, 69(1): 17–28, 1954.

SAS Institute Inc. *SAS/STAT User's Guide, Version 9.3*. SAS Institute, Cary, NC, 1st edition, 2011.

W. Schaafsma and G. N. van Vark. Classification and discrimination problems with applications. part iia. *Satistica Neerlandica*, 33:91–126, 1979.

SonoBat. Call Parameters. `http://www.sonobat.com/SonoBat%20parameters.html`.

Joseph M. Szewczak. A consistent acoustic feature to discriminate myotis species. *Bat Research News*, 41(4):141, 2000.

U.S. Fish and Wildlife Service. Indiana Bat (Myotis sodalis). `http://www.fws.gov/midwest/Endangered/mammals/inba/index.html`, a.

U.S. Fish and Wildlife Service. 2011 Rangewide Population Estimate for the Indiana Bat (Myotis sodalis) by USFWS Region. `http://www.fws.gov/midwest/Endangered/mammals/inba/pdf/2011inbaPopEstimate04Jan12.pdf`, b.

U.S. Fish and Wildlife Service. 2012 Rangewide Summer Guidance IBAT Survey Protocol. `http://www.batprotocol.info/batprotocol.info/2012_Protocol.html`, c.

U.S. Fish and Wildlife Service. White-nose syndrome: A devastating disease of north american bats. `http://www.fws.gov/WhiteNoseSyndrome/pdf/White-nose_fact_sheet_4-2012.pdf`, d.

Kimery C. Vories. Implications of the occurrence and spread of the white-nose syndrome to protection of endangered bats under smcra. In R. I. Barnhisel, editor, *Bridging Reclamation, Science and the Community*, pages 1304–1319, Pittsburgh, PA, 2010. National Meeting of the American Society of Mining and Reclamation.