# Discriminant Analysis With Uncertain Group Labels

Carolyn H. Lichtenstein
Kensington, MD 20895

Steven J. Schwager
Biometrics Unit, Cornell University, Ithaca, NY 14853

BU-1007-MC

September 1992

## Author's Footnote

Carolyn H. Lichtenstein is a statistical consultant, 10905 Drumm Avenue, Kensington, MD 20895. Steven J. Schwager is Associate Professor of Biological Statistics, Biometrics Unit and Statistics Center, Cornell University, Ithaca, NY 14853. The authors would like to acknowledge and thank Larry Brown, Jon Kettenring, Olivia Mitchell, and Paul Velleman for their important contributions to this work.

## ABSTRACT

Traditional discriminant analysis assumes that the group label of each observation is known. However, discriminant analysis may be appropriate even for situations in which the group or classification labels of the cases are not known with certainty. This paper proposes a method for analyzing data of this sort and develops the properties of that method. Maximum likelihood estimates of the parameters under two models of label uncertainty are presented, and their strong consistency and asymptotic normality under suitable assumptions are established. Hypothesis tests are given for whether the original labels are correct, and for which model of uncertainty is more appropriate. Computer implementation of these methods is discussed and an empirical illustration is presented.

The discriminant technique proposed here incorporates a model of the uncertainty in the observations' group labels into the estimation of the discriminant vectors, including a parameter that measures the uncertainty, or error, in the labels. Two models are examined, one assuming a common label error across groups, and one assuming a different label error for each group. The population means and a common covariance matrix are estimated simultaneously with the label error parameters assuming multivariate normal populations. Estimates of the true group labels are provided, as well.

The discriminant vectors are computed similarly to traditional discriminant vectors, from between-groups and within-groups covariance matrices based on the maximum likelihood estimates. The proposed discriminant vectors are shown to be consistent, in contrast to the traditional vectors, which are shown to be asymptotically biased when the labels are uncertain. In addition, the overall error rate of the classification rule based on the maximum likelihood estimates is shown to achieve the Bayes risk in the limit.

The techniques presented in the paper are illustrated for a data set consisting of married women who have been grouped by their labor force status (employed, unemployed, or not in the labor force).

Although the labor force status labels are clearly defined by the government, these definitions are in some sense arbitrary and may not accurately reflect the true groupings among married women. The empirical analysis demonstrates that the new techniques are easy to use and provide interesting results that are quite different from those obtained using traditional discriminant techniques. The new techniques uncover a large amount of uncertainty in the original labels and provide more accurate characterizations of the conceptual groups.

KEY WORDS: Discriminant analysis; Classification; Maximum likelihood estimation.

# 1. THE PROBLEM

Discriminant analysis is concerned with the problem of distinguishing population groups on the basis of observed characteristics. Functions (usually linear) of these characteristics are developed to describe the groups and to classify individuals as members of the groups. However, the population groups must be unambiguously specified before they can be analyzed by these methods.

In some situations, the conceptual population groups can be defined by a single characteristic that can be observed and measured accurately. Samples of observations that have been grouped on the basis of this variable can then be used to estimate the relationships among the population groups as functions of other variables. A good example of this situation is distinguishing, on the basis of various clinical measurements, between people who die from a particular disease and those who do not.

There are many situations, however, in which the conceptual groups are difficult to define quantitatively even when they can be described very clearly qualitatively, often by a single characteristic. The difficulty in the quantitative definition may come from the mechanism used in measuring the particular grouping variable, or it may come from the inability of any single available variable to reflect the groupings accurately. In these cases, sampled observations cannot be assigned definitively to the conceptual groups, so estimating the intergroup relationships is more difficult.

An example of the situation in which the qualitative characteristic cannot be measured accurately is the assignment of pottery shards recovered in an archeological dig to their places of manufacture. Every piece of pottery was made in a particular place, so the conceptual groups are clear, but the incomplete information available to the archeologist thousands of years later does not allow a definitive assignment of pottery to location. An example of the situation in which no single available variable accurately describes the groups is the categorization of women in the U.S. by labor force status (employed, unemployed, or not in the labor force), since these categories were defined by the government in a hierarchical fashion that may not accurately capture the groupings among married women (e.g., these categories might be redefined, with "homemakers" as a new category).

In these cases, the variable that best describes the groupings should be used, e.g., the archeologist's best guess as to the place of manufacture. However, it will not be certain how accurately the group labels assigned to sample observations represent the true conceptual group labels. This, in turn, will lead to uncertainty about whether intergroup relationships estimated on the basis of the sample labels accurately reflect the relationships among the conceptual groups.

The problem of uncertain groupings of observations is more widespread than is generally recognized. Many data sets assumed to involve known labels really involve labels with some ambiguity. In the past, this ambiguity has rarely been recognized, in part because there has been no coherent body of methodology to handle it.

## 2. BACKGROUND

### 2.1. The Literature

Two standard methods could be used in analyzing data of the type described above, although neither is completely satisfactory. One of these methods is traditional discriminant analysis, in which the assigned labels are assumed to be correct. The other method of analysis is clustering, in which the assigned labels are ignored completely. Obviously, neither of these two methods is quite correct: the first assumes information that may not be correct, while the second throws information away.

The approach presented in this paper for data with uncertain labels utilizes discriminant analysis, rather than clustering analysis, as its basis. The literature on traditional discriminant and classification techniques includes complete books (e.g., Hand 1981; Lachenbruch 1975; McLachlan 1992), as well as chapters in most multivariate statistics texts (e.g., Anderson 1984; Kshirsagar 1972; Gnanadesikan 1977).

Relatively little literature is helpful for the situation of data with uncertain group labels. Several papers address the problem of deriving discriminant functions and classification techniques when the observations come from a mixture of two normal distributions and are initially unclassified (Day 1969; Ganesalingam and McLachlan 1978, 1979; Ashikaga and Chang 1981). A few papers present results for the situation when at least some of the observations are mislabeled initially

(Lachenbruch 1966; McLachlan 1972; Lachenbruch 1974). The pattern recognition literature contains several papers closely related to the research presented here, in that they utilize a model of label imperfection in estimating discriminant functions (Shanmugam and Breiphol 1971) and classification probabilities (Chittineni 1982).

### 2.2 Traditional Discriminant Analysis

Assume there are g p-dimensional groups or populations, $\pi_1,\cdots,\pi_g$. Let $q_j$ denote the probability that a randomly chosen observation comes from $\pi_j$. A sample of n observations is taken, $\tilde{n}_j$ of which are labeled as coming from $\pi_j$, so that $\sum_{j=1}^{g} \tilde{n}_j = n$. Let $\tilde{q}_j$ denote the probability that a randomly chosen observation is labeled as coming from $\pi_j$. The observations are $p \times 1$ vectors $\mathbf{x}_{ij}$, $i=1,\cdots,\tilde{n}_j$, $j=1,\cdots,g$. Let $n_j$ denote the number of observations that truly come from $\pi_j$, so that $\sum_{j=1}^{g} n_j = n$ as well. When the labels are uncertain, the $n_j$ are unknown and are part of the information to be estimated from the data.

When the labels are certain, $\tilde{n}_j = n_j$ for all j. In that case, the observations are $\mathbf{x}_{ij}$, $i=1,\cdots,n_j$, $j=1,\cdots,g$, and the sample mean of the $j^{th}$ group is $\bar{\mathbf{x}}_j = (1/n_j) \sum_{i=1}^{n_j} \mathbf{x}_{ij}$. This is the type of data for which traditional discriminant analysis techniques are appropriate. A brief review of these techniques is now presented.

Discriminant analysis has three goals: (1) the determination of which variables (dimensions) are most important in separating the g groups, (2) the classification of new, unlabeled observations, and (3) the extraction of a low-dimensional space of maximal group separation into which the data can be projected for graphical and explanatory purposes. The usual method is based on maximizing the ratio of between-groups to within-groups variation, $F_{\mathbf{a}} = \mathbf{a'Ba/a'Wa}$, with respect to $\mathbf{a}$, where $\mathbf{a}'$ denotes the transpose of the $p \times 1$ vector $\mathbf{a}$, $\mathbf{B} = [1/(g-1)] \sum_{j=1}^{g} n_j(\bar{\mathbf{x}}_j-\bar{\mathbf{x}})(\bar{\mathbf{x}}_j-\bar{\mathbf{x}})'$ is the $p \times p$ sample between-groups covariance matrix, $\mathbf{W} = [1/(n-g)] \sum_{j=1}^{g} \sum_{i=1}^{n_j} (\mathbf{x}_{ij}-\bar{\mathbf{x}}_j)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_j)'$ is the $p \times p$ sample within-groups covariance matrix, and $\bar{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^{g} n_j \bar{\mathbf{x}}_j$ is the $p \times 1$ overall mean vector.

The vector that maximizes $F_{\mathbf{a}}$ is the eigenvector $\mathbf{a}_1$ corresponding to the largest eigenvalue $\ell_1$ of $\mathbf{W}^{-1}\mathbf{B}$. The vector $\mathbf{a}_2$ that produces the next largest F-ratio is the eigenvector corresponding to the

second largest eigenvalue $\ell_2$ of $\mathbf{W}^{-1}\mathbf{B}$, and so on for the succeeding eigenvalues. There will be r = min(p,g–1) eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ corresponding to r nonzero (not necessarily distinct) eigenvalues. To ensure that these eigenvectors are distinct, the $\mathbf{a}_j$ are usually constrained by the condition that $\mathbf{A}_r'\mathbf{W}\mathbf{A}_r = \mathbf{I}$, where $\mathbf{A}_r = [\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_r]$ and $\mathbf{I}$ is the r×r identity matrix. Geometrically, these eigenvectors define the r-dimensional discriminant space in which the projections of the group means are most separated relative to the dispersion of the observations.

The linear combinations $\mathbf{a}'\mathbf{x}$ are commonly known as the discriminant functions. The coefficient of a particular variable in a discriminant function represents the relative importance of that variable to the function. In addition, the correlations of the individual variables with each of the discriminant functions can be of use in evaluating the relative contributions of the variables to the functions.

The usual method for classifying a new observation into one of the g groups is to transform both the new observation and the group means into the discriminant space and then to classify by minimum Euclidean distance. If all r possible eigenvectors are used to define the discriminant space, this procedure is equivalent to classifying the observation by minimum Mahalanobis distance. This classification method assumes that the $q_j$ are equal for all g groups. If, however, estimates of the $q_j$ are available (the usual being $n_j/n$), they can be incorporated into the analysis, leading to classification by the maximum posterior probability of the observation (see Anderson 1984 for further details).

The usual goodness-of-fit criterion of the discriminant procedure, and a measure of the accuracy of the classification rule, is the probability of misclassification. For j,k=1,⋯,g, let $P_{kj} = P(\mathbf{x}$ is classified as $\pi_k|\mathbf{x}$ comes from $\pi_j)$. Thus, there are g(g–1) misclassification probabilities, given by all $P_{kj}$ with k≠j, although the g–1 probabilities $P_{kj}$, k≠j, associated with a particular group j are often combined as $Q_j = P(\mathbf{x}$ is misclassified$|\mathbf{x}$ comes from $\pi_j) = \sum_{k\neq j} P_{kj} = 1 - P_{jj}$. Unfortunately, the expressions for both $P_{kj}$ and $Q_j$ are intractable, making it impossible to calculate the exact misclassification probabilities. There are several estimators for these probabilities. The most widely used, the resubstitution method, tends to underestimate the true misclassification probabilities, but several methods have been proposed for correcting this bias.

The Bayesian approach to discriminant analysis begins with a prior density $p(\boldsymbol{\theta})$ on the vector of parameters $\boldsymbol{\theta}$ for the densities of the g groups. The Bayesian posterior density of $\boldsymbol{\theta}$ given the data **D** can be obtained via $f(\boldsymbol{\theta}|\mathbf{D}) \propto p(\boldsymbol{\theta})f(\mathbf{D}|\boldsymbol{\theta})$, where $f(\mathbf{D}|\boldsymbol{\theta})$ is the conditional density of the data **D** given that the parameter vector equals $\boldsymbol{\theta}$. To find the likelihood of a new observation **x**, conditional on the data **D** and assuming that **x** comes from $\pi_j$, average the density of **x** given $\boldsymbol{\theta}$, $f_j(\mathbf{x}|\boldsymbol{\theta})$, with respect to the posterior density $f(\boldsymbol{\theta}|\mathbf{D})$, obtaining $L_j(\mathbf{x}|\mathbf{D}) = \int f_j(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\mathbf{D})d\boldsymbol{\theta}$. This conditional likelihood $L_j(\mathbf{x}|\mathbf{D})$ is also called the predictive density of **x** within group j. Combining these likelihoods with prior group probabilities $q_j$, j=1,$\cdots$,g, we can estimate the posterior probability that **x** comes from $\pi_j$ as $q_jL_j(\mathbf{x}|\mathbf{D})/\sum_{k=1}^{g} q_kL_k(\mathbf{x}|\mathbf{D})$. For further details and a semi-Bayesian approach to discriminant analysis, see Geisser (1982) and other references given in McLachlan (1992, Sec. 2.2).

## 3. THEORETICAL RESULTS FOR THE UNCERTAIN LABELS PROBLEM

### 3.1. Assumptions and Model

The method presented here incorporates label uncertainty in estimating the discriminant functions and provides estimates of the correct labels. The following assumptions are made:

(1) each observation truly comes from one and only one group, i.e., a unique true label does exist for each observation, even though this label is unknown;

(2) the labeling mechanism does not rely on the data in assigning labels, i.e., the prior label information and the data are independent; and

(3) the groups have multivariate normal distributions with different means but a common covariance matrix $\Sigma$, i.e., $\pi_k \sim \text{MVN}(\boldsymbol{\mu}_k,\Sigma)$ for k=1,$\cdots$,g, and the parameters are unknown.

To develop the discriminant functions, a between-groups covariance matrix is needed. Define the p$\times$p population between-groups covariance matrix to be $\boldsymbol{\Delta}_q = \sum_{j=1}^{g} q_j(\boldsymbol{\mu}_j-\bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_j-\bar{\boldsymbol{\mu}})'$, where the overall population mean is $\bar{\boldsymbol{\mu}} = \sum_{j=1}^{g} q_j\boldsymbol{\mu}_j$ and $q_j$ is the prior probability that a randomly chosen observation comes from $\pi_j$.

The uncertainty associated with each label can be quantified as the probability that the label is correct. If this probability, or some estimate of it obtained independently of the current data, is

available for each observation, it can be incorporated directly into the estimation procedure. However, because this probability will not be available in most data sets, it must be modeled.

Let $\mathbf{x}_{ik}$ denote the $i^{th}$ observation labeled (perhaps erroneously) as k, i.e., as coming from $\pi_k$, for i=1,$\cdots$,$\tilde{n}_k$, k=1,$\cdots$,g. A simple and intuitive model of label uncertainty is:

$$\begin{aligned} P(x_{ik} \in \pi_k) &= P(\mathbf{x}_{ik} \text{ truly comes from } \pi_k) = 1-\epsilon \quad \text{for } i=1,\cdots,\tilde{n}_k,\ k=1,\cdots,g \\ P(x_{ik} \in \pi_j) &= P(\mathbf{x}_{ik} \text{ truly comes from } \pi_j) = \tfrac{\epsilon}{g-1} \quad \text{for } i=1,\cdots,\tilde{n}_k,\ j,k=1,\cdots,g,\ j\neq k \end{aligned} \tag{1}$$

where $0 < \epsilon < 1$. These probabilities are the prior probabilities of group membership for each (labeled) observation. Of course, $\sum_{j=1}^{g} P(\mathbf{x}_{ik} \in \pi_j) = 1$. This model's assumption of a common level of uncertainty for all observations will be generalized to reflect group-dependent error rates in Section 3.5. For the sake of simplicity, we omit explicit conditioning on the value of the uncertain label k in (1) and throughout the rest of the paper; this conditioning is implicit in the subscript of observation $\mathbf{x}_{ik}$.

The parameter $\epsilon$ can be interpreted as the overall level of error in the labeling mechanism. Thus, if $\epsilon = 0$, all the given labels are correct, while if $\epsilon = (g-1)/g$, the given labels have been assigned randomly and contain no information. If $\epsilon = 1$, all the given labels are wrong; the labeling mechanism knows the correct labels and deliberately mislabels every observation randomly. Note that the case of $\epsilon = 1$ does not include the situation in which the observations are clustered correctly but the actual labels associated with the groupings are wrong, e.g., all observations truly from $\pi_j$ are labeled k and vice versa. The latter situation is actually compatible with an $\epsilon$ of 0 since the actual value of the label is not used by the technique (the label is used only to identify the observations' groupings). Although the case of $\epsilon$ equal to 1 may never occur, an $\epsilon$ greater than $(g-1)/g$ might be plausible, as it means that more of the labels were incorrect than would result from random labeling (e.g., if the labeling mechanism has some misconception about the groups).

The density of each observation is a mixture of the normal densities of the g groups, with mixing parameters coming from the model of label uncertainty. Under model (1), the density of observation $\mathbf{x}_{ik}$ is

$$f_k(\mathbf{x}_{ik}) = (1-\epsilon)h_k(\mathbf{x}_{ik}) + \tfrac{\epsilon}{g-1} \sum_{j \neq k} h_j(\mathbf{x}_{ik}) \,, \tag{2}$$

where $h_j(\mathbf{x}_{ik})$ is the MVN($\boldsymbol{\mu}_j$,$\Sigma$) density associated with $\pi_j$. The posterior probabilities of group membership under model (1) are

$$P(k|\mathbf{x}_{ik}) = P(\mathbf{x}_{ik} \in \pi_k|\mathbf{x}_{ik}) = (1-\epsilon)h_k(\mathbf{x}_{ik})/f_k(\mathbf{x}_{ik})$$

$$P(j|\mathbf{x}_{ik}) = P(\mathbf{x}_{ik} \in \pi_j|\mathbf{x}_{ik}) = \tfrac{\epsilon}{g-1} h_j(\mathbf{x}_{ik})/f_k(\mathbf{x}_{ik}) \qquad \text{for } j \neq k \,.$$

Like the prior probabilities in model (1), these posterior probabilities are conditional on the (uncertain) label k of the observation $\mathbf{x}_{ik}$.

### 3.2. Parameter Estimation

The method used most often to estimate the parameters of a normal mixture density is maximum likelihood (see Day 1969; Hosmer 1973; Hassleblad 1966). The maximum likelihood estimators (m.l.e.'s) of the parameters of the densities given in (2), denoted by $\hat{\boldsymbol{\mu}}_j$, $\hat{\Sigma}$, and $\hat{\epsilon}$, are the solution to the equations:

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})\mathbf{x}_{ik}}{\sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})}$$

$$\hat{\Sigma} = \tfrac{1}{n} \sum_{j=1}^{g} \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})(\mathbf{x}_{ik} - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}_{ik} - \hat{\boldsymbol{\mu}}_j)' \tag{3}$$

$$\hat{\epsilon} = \tfrac{1}{n} \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \sum_{j \neq k} \hat{P}(j|\mathbf{x}_{ik}) = 1 - \tfrac{1}{n} \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(k|\mathbf{x}_{ik})$$

where

$$\hat{P}(k|\mathbf{x}_{ik}) = \frac{(1-\hat{\epsilon})\hat{e}_{ikk}}{(1-\hat{\epsilon})\hat{e}_{ikk} + \frac{\hat{\epsilon}}{g-1} \sum_{\ell \neq k} \hat{e}_{ik\ell}}$$

$$\hat{P}(j|\mathbf{x}_{ik}) = \frac{\frac{\hat{\epsilon}}{g-1} \hat{e}_{ikj}}{(1-\hat{\epsilon})\hat{e}_{ikk} + \frac{\hat{\epsilon}}{g-1} \sum_{\ell \neq k} \hat{e}_{ik\ell}} \qquad \text{for } j \neq k \tag{4}$$

and

$$\hat{e}_{ikj} = \exp[-\tfrac{1}{2}(\mathbf{x}_{ik} - \hat{\boldsymbol{\mu}}_j)' \hat{\Sigma}^{-1} (\mathbf{x}_{ik} - \hat{\boldsymbol{\mu}}_j)] \,.$$

These m.l.e.'s must be computed iteratively from given starting values. If the starting value for $\hat{\epsilon}$ is between 0 and 1, the final solution for $\hat{\epsilon}$ is guaranteed to be between 0 and 1 because in that case the $\hat{P}(\cdot|\mathbf{x}_{ik})$ will be between 0 and 1 at each iteration. The properties of the computer algorithm will be discussed further in Section 4. There may be several local maxima or solutions to (3). In principle, one should check explicitly that the global maximum has been achieved and the m.l.e.'s found.

These estimators are intuitively sensible. The terms $\hat{P}(j|\mathbf{x}_{ik})$ are estimators of the posterior probabilities of membership defined in Section 3.1. Any particular $\hat{\boldsymbol{\mu}}_j$ at a given step in the iterative process is a weighted average of all the observations, with the weights being equal to the posterior probabilities, estimated in that step, that the observations belong to $\pi_j$. The quantity $\hat{\boldsymbol{\Sigma}}$ can be interpreted as a weighted covariance matrix. The estimator of $\epsilon$ is the average, over all the observations, of the $\sum_{j \neq k} \hat{P}(j|\mathbf{x}_{ik})$ (the total posterior probability that the label of $\mathbf{x}_{ik}$ is wrong); $\hat{\epsilon}$ can be interpreted as the estimated overall level of error in the original labels.

The estimated between-groups covariance matrix, $\hat{\boldsymbol{\Delta}}_{\hat{\boldsymbol{q}}}$, is defined as

$$\hat{\boldsymbol{\Delta}}_{\hat{\boldsymbol{q}}} = \tfrac{1}{\tilde{n}}\,\mathbf{T} - \hat{\boldsymbol{\Sigma}} = \tfrac{1}{\tilde{n}} \sum_{j=1}^{g} \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})(\hat{\boldsymbol{\mu}}_j - \hat{\tilde{\boldsymbol{\mu}}})(\hat{\boldsymbol{\mu}}_j - \hat{\tilde{\boldsymbol{\mu}}})'$$

where $\mathbf{T} = \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} (\mathbf{x}_{ik} - \bar{\mathbf{x}})(\mathbf{x}_{ik} - \bar{\mathbf{x}})'$ is the total sums of squares and cross-products matrix and $\hat{\tilde{\boldsymbol{\mu}}}$ $= (1/n) \sum_{j=1}^{g} \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})\hat{\boldsymbol{\mu}}_j$ is the estimated overall population mean. Comparing $\hat{\boldsymbol{\Delta}}_{\hat{\boldsymbol{q}}}$ to the population matrix $\boldsymbol{\Delta}_{\boldsymbol{q}}$ suggests that $\hat{n}_j = \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})$ can be considered an estimate of $n_j$ (which is unknown when the labels are uncertain), so that $\hat{q}_j = \hat{n}_j/n$ is an estimate of $q_j$, the true probability associated with group $\pi_j$.

The estimated discriminant vectors are defined as the eigenvectors $\hat{\boldsymbol{\alpha}}_1, \cdots, \hat{\boldsymbol{\alpha}}_r$ of $\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Delta}}_{\hat{\boldsymbol{q}}}$ that satisfy $\hat{\mathbf{A}}_r'\hat{\boldsymbol{\Sigma}}\hat{\mathbf{A}}_r = \mathbf{I}$ where $\hat{\mathbf{A}}_r = [\hat{\boldsymbol{\alpha}}_1, \cdots, \hat{\boldsymbol{\alpha}}_r]$. As in the traditional case, $r = \min(p, g-1)$ when the $\hat{\boldsymbol{\mu}}_j$ are linearly independent, and the rank of $\hat{\boldsymbol{\Delta}}_{\hat{\boldsymbol{q}}}$ is r. The interpretation of the coefficients in the discriminant functions, $\hat{\boldsymbol{\alpha}}_j'\mathbf{x}$, is the same as the interpretation of the coefficients of the traditional discriminant functions; they give the relative weights of the variables in separating the groups. Of course, just as in the traditional case, it is possible to choose a subset of the eigenvectors to define the

discriminant space. If the eigenvalues become very small at some point, i.e., $\hat{\lambda}_{s+1},\cdots,\hat{\lambda}_r$ are very much smaller than $\hat{\lambda}_1,\cdots,\hat{\lambda}_s$, then one might want to conclude that the population discriminant space had only s dimensions and only $\hat{\alpha}_1,\cdots,\hat{\alpha}_s$ were necessary. Unfortunately, it is difficult to formulate an exact hypothesis test of the significance of the eigenvalues because of the unknown nature of the distribution of $\hat{\Sigma}$.

The final parameter to be estimated is the true label of each observation. This label can be thought of as a vector, as follows. Let $\mathbf{v}_j$ denote the $1 \times g$ vector $(0,\cdots,0,1,0,\cdots,0)$ with the $j^{th}$ element equal to 1 and all other elements equal to 0. A true label vector $\mathbf{L}_{ik}$ of length g can be defined for an observation $\mathbf{x}_{ik}$ that truly comes from $\pi_j$ as $\mathbf{L}_{ik} = \mathbf{v}_j$. The vector of sample posterior probabilities associated with that observation, $\left(\hat{P}(1|\mathbf{x}_{ik}),\cdots,\hat{P}(g|\mathbf{x}_{ik})\right)$, is an estimate of the true label vector of $\mathbf{x}_{ik}$. In fact, it is the posterior Bayes estimator of $\mathbf{L}_{ik}$ with respect to the prior distribution composed of the estimated probabilities $\hat{P}(\mathbf{L}_{ik} = \mathbf{v}_j)$. This can be seen from the following discussion. Under the model of label uncertainty presented in (1), the estimated prior distribution of each $\mathbf{L}_{ik}$ is multinomial with n = 1, $\hat{P}(\mathbf{L}_{ik} = \mathbf{v}_k) = 1-\hat{\epsilon}$, and $\hat{P}(\mathbf{L}_{ik} = \mathbf{v}_j) = \frac{\hat{\epsilon}}{g-1}$ for each $j \neq k$. The estimated posterior distribution of $\mathbf{L}_{ik}$ reduces to the expressions defined in (4),

$$P(\mathbf{L}_{ik} = \mathbf{v}_j|\mathbf{x}_{ik}) = \begin{cases} \hat{P}(k|\mathbf{x}_{ik}) & \text{for } j=k \\ \hat{P}(j|\mathbf{x}_{ik}) & \text{for } j\neq k \end{cases}.$$

Since the mean of this distribution is the posterior Bayes estimator of the label vector, we get

$$\begin{aligned} \hat{\mathbf{L}}_{ik} &= \textstyle\sum_{j=1}^{g} \mathbf{v}_j \hat{P}(j|\mathbf{x}_{ik}) \\ &= \left(\hat{P}(1|\mathbf{x}_{ik}),\cdots,\hat{P}(g|\mathbf{x}_{ik})\right). \end{aligned} \tag{5}$$

When a classification decision is not the sole objective, the analyst may obtain additional useful information from the relative magnitudes of the probabilities in $\hat{\mathbf{L}}_{ik}$.

When one is interested in classifying a new, unlabeled observation, $\mathbf{x}_0$, to a group, the vector of estimated posterior probabilities,

$$\hat{P}(j|\mathbf{x}_0) = \frac{\hat{q}_j \exp\left[-\frac{1}{2}(\mathbf{x}_0-\hat{\boldsymbol{\mu}}_j)'\hat{\Sigma}^{-1}(\mathbf{x}_0-\hat{\boldsymbol{\mu}}_j)\right]}{\sum_{k=1}^{g} \hat{q}_k \exp\left[-\frac{1}{2}(\mathbf{x}_0-\hat{\boldsymbol{\mu}}_k)'\hat{\Sigma}^{-1}(\mathbf{x}_0-\hat{\boldsymbol{\mu}}_k)\right]},$$

can be used. $\hat{P}(j|\mathbf{x}_0)$ is based on less information than $\hat{P}(j|\mathbf{x}_{ik})$ because it is not conditional on an uncertain label as $\hat{P}(j|\mathbf{x}_{ik})$ is. Of course, if the new observation has been assigned an uncertain label, the vector given in (5) can be computed instead.

It is also possible to make a classification decision for a new, unlabeled observation by extending the traditional discriminant analysis methods to the uncertain labels model. One choice of decision rule is to classify $\mathbf{x}_0$ as $\pi_j$ if $\hat{P}(j|\mathbf{x}_0) = \max_k \hat{P}(k|\mathbf{x}_0)$. This is equivalent to the rule of classifying $\mathbf{x}_0$ as $\pi_j$ if $\hat{\delta}^2_{0j} - 2\log \hat{q}_j = \min_k (\hat{\delta}^2_{0k} - 2\log \hat{q}_k)$ where $\hat{\delta}^2_{0k} = (\mathbf{x}_0 - \hat{\boldsymbol{\mu}}_k)' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_0 - \hat{\boldsymbol{\mu}}_k)$ is the estimated Mahalanobis distance of $\mathbf{x}_0$ from $\hat{\boldsymbol{\mu}}_k$. This rule can be restated as classifying $\mathbf{x}_0$ as $\pi_j$ if

$$(\mathbf{z}_0 - \hat{\boldsymbol{\nu}}_j)'(\mathbf{z}_0 - \hat{\boldsymbol{\nu}}_j) - 2\log \hat{q}_j = \min_k \left[ (\mathbf{z}_0 - \hat{\boldsymbol{\nu}}_k)'(\mathbf{z}_0 - \hat{\boldsymbol{\nu}}_k) - 2\log \hat{q}_k \right],$$

where $\mathbf{z}_0 = \hat{\mathbf{A}}_r' \mathbf{x}_0$, $\hat{\boldsymbol{\nu}}_k = \hat{\mathbf{A}}_r' \hat{\boldsymbol{\mu}}_k$, and $\hat{\mathbf{A}}_r = [\hat{\boldsymbol{\alpha}}_1, \cdots, \hat{\boldsymbol{\alpha}}_r]$ so that $(\mathbf{z}_0 - \hat{\boldsymbol{\nu}}_k)'(\mathbf{z}_0 - \hat{\boldsymbol{\nu}}_k)$ is the Euclidean distance of $\mathbf{x}_0$ from $\hat{\boldsymbol{\mu}}_k$ in the discriminant space. Recall that this equivalence holds only if all r eigenvectors corresponding to the r nonzero eigenvalues are used. When the $\hat{q}_j$ are all equal, this classification method is based on minimizing the estimated Mahalanobis distance of $\mathbf{x}_0$ from $\hat{\boldsymbol{\mu}}_k$, but when the $\hat{q}_j$ are unequal, the classification is weighted towards those groups with higher prior probabilities.

These classification methods can also be used to "classify" each original observation to obtain an estimate of the true label that is different from the estimate given by (5). The only modification is that $\hat{P}(\cdot|\mathbf{x}_{ik})$ as defined in the m.l.e. algorithm is used in place of $\hat{P}(\cdot|\mathbf{x}_0)$. This procedure is equivalent to using only the largest element of the vector given in (5) instead of the whole vector. Since it is often desirable to be able to assign each observation to a specific group, this procedure may be useful, although it does throw away the information present in the actual values of the conditional probabilities. Thus, combining the two labels may be the best solution, i.e., by making a classification decision but also examining the vector given in (5).

The true conditional probability of misclassification of the estimated classification rule given above is $P_j = P\left[\hat{P}(j|\mathbf{x}) \neq \max_m \hat{P}(m|\mathbf{x}) \mid \mathbf{x} \in \pi_j\right]$. Unfortunately, the intractability of the distribution of the $\hat{P}(\cdot|\mathbf{x})$ prevents the exact evaluation of these misclassification probabilities. The true overall conditional probability of misclassification associated with the optimal classification regions for

multivariate normal groups can be defined as $P_M = \sum_{j=1}^{g} q_j \sum_{k \neq j} P_{kj} = \sum_{j=1}^{g} q_j Q_j$, where $P_{kj}$ and $Q_j$ are defined in Section 2.2. This is the minimum error rate attainable. Again, it is impossible to evaluate analytically the multiple integrals needed to derive the value of $P_M$.

The individual conditional misclassification probabilities defined above cannot be estimated from the sample because the observations' true labels are unknown. However, it is possible to estimate the overall error rate $P_M$. The estimator suggested here is $\hat{E}_0 = 1 - \frac{1}{n_0} \sum_{i=1}^{n_0} \max_m \hat{P}(m|\mathbf{x}_i)$, where the $\mathbf{x}_i$ are a sample of $n_0$ unlabeled observations to be classified that are independent of the sample used to compute the m.l.e.'s.

Ganesalingam and McLachlan (1980) examined the properties of $\hat{E}_0$ for the two-group case. First they derived the asymptotic bias of $\hat{E}_0$ in estimating $P_M$, which equals $q_1P_1 + q_2P_2$ in the two-group case. They then performed computer simulations to evaluate the bias of $\hat{E}_0$ when n is small. Results were derived assuming three different distributions of the observations, one of which was a normal mixture. Ganesalingam and McLachlan concluded, for the normal mixture case, that $\hat{E}_0$ performs similarly to the "plug-in" estimator, which is used frequently in traditional discriminant analysis and is designed specifically for normal groups with a common covariance matrix.

If it is impossible to obtain an independent sample of $n_0$ observations, the estimator

$$\hat{E} = 1 - \frac{1}{n} \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \max_m \hat{P}(m|\mathbf{x}_{ik}) \tag{6}$$

can be used, where now the posterior probabilities are associated with the same observations that were used in computing the m.l.e.'s. Of course, this estimator will be biased downward, similarly to the resubstitution estimator discussed in connection with traditional discriminant analysis.

In deriving the maximum likelihood results, we have assumed that $\epsilon$ is completely unknown and must be estimated from the likelihood equations. It is possible, however, that $\epsilon$ either is known or can be accurately estimated from previous data, e.g., if the labels were assigned by a machine for which the labeling error was known from the method of construction of the machine or could be estimated from previous test runs of the machine. In this case, the m.l.e.'s of the $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}$ are those given in (3) with $\hat{\epsilon}$ replaced by this prior value of $\epsilon$; there is, of course, no expression for $\hat{\epsilon}$.

A special case of the situation in which $\epsilon$ is known is $\epsilon = 0$. This is equivalent to having labels that are certain, and traditional discriminant techniques are appropriate. In fact, the maximum likelihood estimates $\hat{\mu}_j$ and $\hat{\Sigma}$ reduce to the traditional estimates, and the discriminant vectors reduce to the traditional vectors.

### 3.3. Asymptotic Properties of the Estimators

Small sample properties of the m.l.e.'s, such as expected values and variances, seem to be impossible to derive theoretically because of the iterative nature of the estimation procedure. However, asymptotic results can be developed. The proofs for all of the following theorems and corollaries are given in the Appendix; full details can be found in Lichtenstein (1985).

*Theorem 1.* The m.l.e.'s as defined in (3) exist in probability and are strongly consistent estimators of the true parameters for $0 \leq \epsilon \leq 1$.

Before presenting the results concerning the asymptotic normality of the m.l.e.'s, we must define the matrix $\mathbf{J} = \left[\mathrm{J}_{rs}(\boldsymbol{\theta})\right]$ whose (r,s)$^{\underline{th}}$ element is given by

$$\mathrm{J}_{rs}(\boldsymbol{\theta}) = \sum_{k=1}^{g} \tilde{\mathrm{q}}_k \int_{R_k} \left(\frac{\partial \ell \mathrm{n}\, \mathrm{f}_k}{\partial \theta_r}\right)\left(\frac{\partial \ell \mathrm{n}\, \mathrm{f}_k}{\partial \theta_s}\right) \mathrm{f}_k \, \mathrm{d}\mathbf{x}_k \, ,$$

where $\mathrm{R}_k$ is the region over which the values of $\mathbf{X}_k$ are defined; $\boldsymbol{\theta}$ is a parameter vector of dimension $\mathrm{pg} + \frac{1}{2}\mathrm{p}(\mathrm{p}+1) + 1$ equal to $(\boldsymbol{\mu}_1', \cdots, \boldsymbol{\mu}_g', \Sigma_{11}, \Sigma_{12}, \Sigma_{22}, \cdots, \Sigma_{pp}, \epsilon)$; $\mathrm{f}_k = \mathrm{f}_k(\mathbf{x}_k)$ is the normal mixture density given in (2); and $\tilde{\mathrm{q}}_k$ is the prior probability of an observation being labeled as group k.

*Theorem 2(i).* The m.l.e.'s $\hat{\boldsymbol{\mu}}_1, \cdots, \hat{\boldsymbol{\mu}}_g$, $\hat{\Sigma}$, and $\hat{\epsilon}$ are asymptotically jointly normally distributed about the true parameter values with covariance matrix $\mathbf{J}_0^{-1}$, where $\mathbf{J}_0 = \left[\mathrm{J}_{rs}(\boldsymbol{\theta}^\circ)\right]$ and $\boldsymbol{\theta}^\circ$ is the vector of true parameter values, for $0 < \epsilon < 1$.

*Theorem 2(ii).* When $\epsilon = 0$, as $\mathrm{n} \to \infty$ the m.l.e.'s $\hat{\boldsymbol{\mu}}_1, \cdots, \hat{\boldsymbol{\mu}}_g$, $\hat{\Sigma}$, and $\hat{\epsilon}$ have the following joint asymptotic distribution, which is a "truncated" normal distribution:

(1) with probability $\frac{1}{2}$, $\hat{\epsilon} = 0$ and $\sqrt{\mathrm{n}}(\hat{\boldsymbol{\mu}}_1, \cdots, \hat{\boldsymbol{\mu}}_g, \hat{\Sigma})$ have a joint normal distribution centered at the true values and with covariance matrix $(\mathbf{J}_0^*)^{-1}$, where $\mathbf{J}_0^*$ is the upper left square submatrix of $\mathbf{J}_0$ of dimension $\mathrm{pg} + \frac{1}{2}\mathrm{p}(\mathrm{p}+1)$; and

(2) with probability $\frac{1}{2}$, $\sqrt{n}(\hat{\mu}_1,\cdots,\hat{\mu}_g,\hat{\Sigma},\hat{\epsilon})$ have a joint normal distribution centered at the true values, with covariance matrix $\mathbf{J}_0^{-1}$ $\left(\text{of dimension } pg + \frac{1}{2}p(p+1) + 1\right)$, and restricted to $\hat{\epsilon} > 0$.

The final asymptotic results follow directly from the consistency of the m.l.e.'s and certain algebraic identities:

(1) The matrix $\hat{\Delta}_{\hat{\mathbf{q}}}$ is consistent, i.e., $\hat{\Delta}_{\hat{\mathbf{q}}} \to \Delta_{\mathbf{q}}$ in probability as $n \to \infty$, $n_j/n \to q_j$ for all j, and $\tilde{n}_k/n \to \tilde{q}_k$ for all k.

(2) The quantity $\hat{q}_j = \hat{n}_j/n = (1/n) \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})$ is a consistent estimator of $q_j$, i.e., $\hat{q}_j \to q_j$ in probability as $n \to \infty$, $n_j/n \to q_j$ for all j, and $\tilde{n}_k/n \to \tilde{q}_k$ for all k.

*Corollary.* The set of eigenvectors $\hat{\boldsymbol{\alpha}}_j$ is a consistent estimator of the set of population eigenvectors $\boldsymbol{\alpha}_j$ as long as the corresponding nonzero eigenvalues are distinct.

The parameter $\epsilon$ deserves attention beyond its role in the estimation of the discriminant vectors and group labels since, as mentioned earlier, $\epsilon$ measures the overall error rate in the labels. It is therefore useful to derive hypothesis tests concerning $\epsilon$. The most interesting hypothesis is that the original labels are correct, i.e., $H_0$: $\epsilon = 0$. A test can be constructed based on the generalized likelihood ratio.

Begin by defining the sample within-groups covariance matrix as $\tilde{\mathbf{W}} = [1/(n-g)] \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} (\mathbf{x}_{ik}-\bar{\mathbf{x}}_k)(\mathbf{x}_{ik}-\bar{\mathbf{x}}_k)'$ and the sample between-groups covariance matrix as $\tilde{\mathbf{B}} = [1/(g-1)] \sum_{k=1}^{g} \tilde{n}_k (\bar{\mathbf{x}}_k-\bar{\mathbf{x}})(\bar{\mathbf{x}}_k-\bar{\mathbf{x}})'$, where $\bar{\mathbf{x}}_k = (1/\tilde{n}_k) \sum_{i=1}^{\tilde{n}_k} \mathbf{x}_{ik}$ and $\bar{\mathbf{x}} = (1/n) \sum_{k=1}^{g} \tilde{n}_k \bar{\mathbf{x}}_k = (1/n) \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \mathbf{x}_{ik}$. The relationship to the matrices $\mathbf{W}$ and $\mathbf{B}$ defined in Section 2.2 is clear.

*Theorem 3.* The $\alpha$-level asymptotic likelihood ratio test of $H_0$: $\epsilon = 0$ rejects $H_0$ when $-2\,\ell n\,\lambda > \chi_1^2(1-2\alpha)$ where

$$\lambda = \frac{|\tilde{\mathbf{W}}|^{-n/2} \prod_{k=1}^{g} \prod_{i=1}^{\tilde{n}_k} \bar{e}_{ikk}}{|\hat{\Sigma}|^{-n/2} \prod_{k=1}^{g} \prod_{i=1}^{\tilde{n}_k} \left[(1-\hat{\epsilon})\hat{e}_{ikk} + \frac{\hat{\epsilon}}{g-1} \sum_{\ell \neq k} \hat{e}_{ik\ell}\right]},$$

$\bar{e}_{ikk} = \exp\left[-\frac{1}{2}(\mathbf{x}_{ik}-\bar{\mathbf{x}}_k)'\tilde{\mathbf{W}}^{-1}(\mathbf{x}_{ik}-\bar{\mathbf{x}}_k)\right]$, the denominator of $\lambda$ is computed from the original m.l.e.'s of (3), and $\chi_1^2(1-2\alpha)$ is the $(1-2\alpha)^{\underline{th}}$ quantile of the $\chi^2$ distribution with 1 degree of freedom.

Of course, it is also possible to construct a likelihood ratio test similar to the one above for a general hypothesis $H_0$: $\epsilon = \epsilon^*$. The form of this asymptotic test is now to reject $H_0$ when $-2\ \ell n\ \lambda > \chi_1^2(1-\alpha)$ where

$$\lambda = \frac{|\hat{\Sigma}_{\epsilon^*}|^{-n/2} \prod_k \prod_i \left[(1-\epsilon^*)\hat{e}^*_{ikk} + \frac{\epsilon^*}{g-1} \sum_{\ell \neq k} \hat{e}^*_{ik\ell}\right]}{|\hat{\Sigma}|^{-n/2} \prod_k \prod_i \left[(1-\hat{\epsilon})\hat{e}_{ikk} + \frac{\hat{\epsilon}}{g-1} \sum_{\ell \neq k} \hat{e}_{ik\ell}\right]}$$

and $\hat{\Sigma}_{\epsilon^*}$ and $\hat{e}^*_{ik\ell}$ are the m.l.e.'s computed with $\epsilon = \epsilon^*$ fixed (see Kendall and Stuart 1979, pp. 246-247).

### 3.4 Assessment of the Estimators

In investigating the merits of $\hat{\alpha}_1'\mathbf{x},\cdots,\hat{\alpha}_r'\mathbf{x}$ as the discriminant functions, we can compare them to the obvious competitors, $\tilde{\alpha}_1'\mathbf{x},\cdots,\tilde{\alpha}_r'\mathbf{x}$, where the $\tilde{\alpha}_j$ are the eigenvectors obtained by using the traditional discriminant technique as if the labels were correct. An important question is whether using traditional techniques on data with uncertain labels, i.e., ignoring the uncertainty, gives much less accurate results than the theoretically correct but more complicated m.l.e. technique. The large-sample properties of the techniques will be compared here.

The limits of $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{B}}$, the sample within-groups and between-groups covariance matrices defined in Section 3.3, can be derived very easily since each $\bar{\mathbf{x}}_k$ is based on $\tilde{n}_k$ i.i.d. observations.

*Theorem 4.*

$$[(n-g)/n]\ \tilde{\mathbf{W}} \rightarrow \mathbf{\Sigma} + \tilde{\mathbf{U}}_\epsilon \qquad \text{with probability 1}$$

and

$$[(g-1)/n]\ \tilde{\mathbf{B}} \rightarrow \mathbf{\Delta}_q - \tilde{\mathbf{U}}_\epsilon \qquad \text{with probability 1 ,}$$

as $n \rightarrow \infty$ and $\tilde{n}_k/n \rightarrow \tilde{q}_k$ for all k, where $\tilde{\mathbf{U}}_\epsilon = \sum_{j=1}^g q_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j' - \sum_{k=1}^g \tilde{q}_k E(\mathbf{x}_k)E(\mathbf{x}_k)'$ is a nonnegative definite matrix, and $E(\mathbf{x}_k) = (1-\epsilon)\boldsymbol{\mu}_k + [\epsilon/(g-1)]\sum_{\ell \neq k} \boldsymbol{\mu}_\ell$, which follows from equation (2).

Consequently, even in the limit, $\tilde{\mathbf{W}}$ will "overestimate" the true within-groups covariance matrix and $\tilde{\mathbf{B}}$ will "underestimate" the true between-groups covariance matrix. The direction of the bias of $\tilde{\mathbf{W}}^{-1}\tilde{\mathbf{B}}$ in estimating $\mathbf{\Sigma}^{-1}\mathbf{\Delta}_q$ is uncertain, since

$$\begin{aligned} \tfrac{g-1}{n-g}\,\tilde{\mathbf{W}}^{-1}\tilde{\mathbf{B}} \;&\rightarrow\; (\boldsymbol{\Sigma}+\tilde{\mathbf{U}}_{\epsilon})^{-1}(\boldsymbol{\Delta}_{q}-\tilde{\mathbf{U}}_{\epsilon}) \\ &=\; (\mathbf{I}+\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{U}}_{\epsilon})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{q}-(\mathbf{I}+\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{U}}_{\epsilon})^{-1}\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{U}}_{\epsilon}\,, \end{aligned} \tag{7}$$

so the bias depends on the actual values of $\boldsymbol{\Sigma}$ and $\tilde{\mathbf{U}}_{\epsilon}$. Because we are ultimately interested only in the consistency of the eigenvectors of $\tilde{\mathbf{W}}^{-1}\tilde{\mathbf{B}}$ in estimating the eigenvectors of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{q}$, it is more appropriate to look at the asymptotic bias of $\tilde{\mathbf{W}}^{-1}\tilde{\mathbf{B}}$ in estimating $a\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{q} + b\mathbf{I}$ (which has the same eigenvectors as $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{q}$). Examining (7) reveals that the bias will be zero when $\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{U}}_{\epsilon} = d\mathbf{I}$ for any constant $d \neq 0$. This is a restrictive condition; many reasonable configurations of the groups would result in $\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{U}}_{\epsilon} \neq d\mathbf{I}$ and a traditional estimate that is asymptotically biased.

We have just shown that using traditional discriminant techniques when the labels are uncertain results in inaccurate estimates; the magnitude of the inaccuracy depends on the configuration of the groups. We now move on to examine how well the correct m.l.e. technique estimates the true parameters.

*Theorem 5.* The estimated error rate given in (6) achieves the Bayes risk in the limit, i.e., $\hat{E} \rightarrow \tilde{R}$ in probability, as $n \rightarrow \infty$ and $\tilde{n}_k/n \rightarrow \tilde{q}_k$ for all k, where $\tilde{R} = \sum_{k=1}^{g} \tilde{q}_k E_{\boldsymbol{x}_k}[\tilde{r}(\mathbf{x}_k)]$ is the Bayes risk and $\tilde{r}(\mathbf{x}_k) = 1 - \max_m P(m|\mathbf{x}_k)$.

### 3.5. Group-Dependent Errors

A model of the uncertainty in the labels that may be more appropriate for many data sets allows for group-dependent errors. Formally, this model is:

$$\begin{aligned} P(\mathbf{x}_{ik} \in \pi_k) &= P(\mathbf{x}_{ik} \text{ truly comes from } \pi_k) = 1 - \epsilon_k \\ P(\mathbf{x}_{ik} \in \pi_j) &= P(\mathbf{x}_{ik} \text{ truly comes from } \pi_j) = \frac{\epsilon_k}{g-1} \qquad \text{for } j \neq k\,, \end{aligned}$$

where $0 \leq \epsilon_k \leq 1$, $k = 1,\cdots,g$.

It is possible to interpret $\epsilon_k$ as the level of error in the $k^{\underline{th}}$ label, i.e., as the error rate of the labeling mechanism in assigning the label k to observations. For example, in the data on the labor force status of U.S. women discussed in Section 1, it is quite possible that the official government category of "not in the labor force" is more accurate than the other two categories because these

women are more homogeneous, e.g., in staying home to take care of small children. Particular values of $\epsilon_k$ can be interpreted similarly to the corresponding values of $\epsilon$. Thus, $\epsilon_k = 0$ means that all the observations labeled k are correctly labeled, while $\epsilon_k = 1$ means that all those observations are labeled incorrectly. And $\epsilon_k = (g-1)/g$ means that the label k was assigned randomly to observations that were equally likely to have come from $\pi_1,\cdots,\pi_g$.

The distribution of an observation $\mathbf{x}$ labeled k is now $f_k(\mathbf{x}) = (1-\epsilon_k)h_k(\mathbf{x}) + [\epsilon_k/(g-1)] \sum_{j\neq k} h_j(\mathbf{x})$, where $h_j(\mathbf{x})$ is the MVN($\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}$) density associated with $\pi_j$. The m.l.e.'s of the parameters are denoted by $\hat{\boldsymbol{\mu}}_j$, $\hat{\boldsymbol{\Sigma}}$, and $\hat{\epsilon}_k$, and are the solution to the equations:

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})\mathbf{x}_{ik}}{\sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})} \qquad \text{for } j = 1,\cdots,g$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{\tilde{n}} \sum_{j=1}^{g} \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})(\mathbf{x}_{ik}-\hat{\boldsymbol{\mu}}_j)(\mathbf{x}_{ik}-\hat{\boldsymbol{\mu}}_j)' \tag{8}$$

$$\hat{\epsilon}_k = \frac{1}{\tilde{n}_k} \sum_{i=1}^{\tilde{n}_k} \sum_{j\neq k} \hat{P}(j|\mathbf{x}_{ik}) = 1 - \frac{1}{\tilde{n}_k} \sum_{i=1}^{\tilde{n}_k} \hat{P}(k|\mathbf{x}_{ik}) \qquad \text{for } k = 1,\cdots,g$$

where

$$\hat{P}(k|\mathbf{x}_{ik}) = \frac{(1-\hat{\epsilon}_k)\hat{e}_{ikk}}{(1-\hat{\epsilon}_k)\hat{e}_{ikk} + \frac{\hat{\epsilon}_k}{g-1} \sum_{\ell\neq k} \hat{e}_{ik\ell}}$$

$$\hat{P}(j|\mathbf{x}_{ik}) = \frac{\frac{\hat{\epsilon}_k}{g-1}\hat{e}_{ikj}}{(1-\hat{\epsilon}_k)\hat{e}_{ikk} + \frac{\hat{\epsilon}_k}{g-1} \sum_{\ell\neq k} \hat{e}_{ik\ell}} \qquad \text{for } j \neq k$$

and

$$\hat{e}_{ikj} = \exp\left[-\tfrac{1}{2}(\mathbf{x}_{ik}-\hat{\boldsymbol{\mu}}_j)'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_{ik}-\hat{\boldsymbol{\mu}}_j)\right].$$

These m.l.e.'s can be interpreted in much the same way as those of Section 3.2, and their asymptotic properties, as well as the proofs of these properties, are essentially the same as those presented in Section 3.3.

The estimated discriminant vectors are then derived in the same way as for the previous model, i.e., $\hat{\boldsymbol{\alpha}}_1,\cdots,\hat{\boldsymbol{\alpha}}_r$ are the eigenvectors of $\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Delta}}_{\hat{q}}$. It is interesting to note that, if $g = 2$, the asymptotic

discriminant function is equivalent to $D_M$ of Lachenbruch (1966). Just as in the common error model, $\hat{\Sigma}^{-1}\hat{\Delta}_{\hat{q}} \rightarrow \Sigma^{-1}\Delta_q$ in probability and $\hat{\alpha}_j \rightarrow \alpha_j$ in probability for $j = 1,\cdots,r$ as $n \rightarrow \infty$ and $\tilde{n}_k/n \rightarrow \tilde{q}_k$ for all k.

As in the common $\epsilon$ case, the sample matrices $\tilde{W}$ and $\tilde{B}$ do not estimate the population matrices correctly. The result of (7) in Section 3.4 holds with the nonnegative definite matrix $\tilde{U}_\epsilon$ defined in Theorem 4 replaced by an analogous but more general matrix $\tilde{U}_{\epsilon_k}$, in which $\epsilon_k$ replaces $\epsilon$ in $E(\mathbf{x}_k)$.

Finally, it is useful to derive hypothesis tests concerning the error terms $\epsilon_k$. There are two hypotheses of major interest in the group-dependent case. The first is the hypothesis that the observations originally labeled k are correctly labeled, i.e., $H_0$: $\epsilon_k = 0$ for a specific, arbitrary k. The second hypothesis is that the error rates in the labels are the same across groups, i.e., $H_0$: $\epsilon_1 = \cdots = \epsilon_g = \epsilon \neq 0$. Both of these hypotheses can be tested using a generalized likelihood ratio.

*Theorem 6(i).* The $\alpha$-level asymptotic likelihood ratio test of $H_0$: $\epsilon_k = 0$ for a single specified k $(1 \leq k \leq g)$ rejects $H_0$ when $-2\ \ell n\ \lambda > \chi^2_1(1-2\alpha)$ where

$$\lambda = \frac{|\hat{\Sigma}_*|^{-n/2} \prod_{h=1}^{\tilde{n}_k} \hat{e}^*_{hkk} \prod_{j\neq k} \prod_{i=1}^{\tilde{n}_j} \left[(1-\hat{\epsilon}^*_j)\hat{e}^*_{ijj} + \frac{\hat{\epsilon}^*_j}{g-1} \sum_{\ell\neq j} \hat{e}^*_{ij\ell}\right]}{|\hat{\Sigma}_{\epsilon_k}|^{-n/2} \prod_{m=1}^{g} \prod_{i=1}^{n_m} \left[(1-\hat{\epsilon}_m)\hat{e}_{imm} + \frac{\hat{\epsilon}_m}{g-1} \sum_{\ell\neq m} \hat{e}_{im\ell}\right]},$$

with the m.l.e.'s in the numerator computed with the particular $\epsilon_k$ set to 0, and with the denominator computed from the original m.l.e.'s in (8).

*Theorem 6(ii).* The $\alpha$-level asymptotic likelihood ratio test of $H_0$: $\epsilon_1 = \cdots = \epsilon_g = \epsilon \neq 0$ rejects $H_0$ when $-2\ \ell n\ \lambda > \chi^2_{g-1}(1-\alpha)$ where

$$\lambda = \frac{|\hat{\Sigma}_\epsilon|^{-n/2} \prod_{k=1}^{g} \prod_{i=1}^{\tilde{n}_k} \left[(1-\hat{\epsilon})\hat{e}^{(\epsilon)}_{ikk} + \frac{\hat{\epsilon}}{g-1} \sum_{\ell\neq k} \hat{e}^{(\epsilon)}_{ik\ell}\right]}{|\hat{\Sigma}_{\epsilon_k}|^{-n/2} \prod_{k=1}^{g} \prod_{i=1}^{\tilde{n}_k} \left[(1-\hat{\epsilon}_k)\hat{e}_{ikk} + \frac{\hat{\epsilon}_k}{g-1} \sum_{\ell\neq k} \hat{e}_{ik\ell}\right]},$$

with the numerator computed assuming a common $\epsilon$ (and thus the m.l.e.'s given by (3)), and with the denominator computed from the m.l.e.'s in (8).

## 4. COMPUTER ALGORITHM

The properties of the simple normal mixture algorithm have been studied fairly extensively (Day 1969; Hassleblad 1966; Everitt and Hand 1980). Since the m.l.e. algorithm presented in this paper is very similar to the algorithm for a simple normal mixture, it is reasonable to assume that many of its properties are also very similar.

There are two different types of methods that are typically used in estimating the parameters of a normal mixture: (1) a special root-finding technique such as Newton-Raphson, and (2) the EM algorithm (Dempster, Laird and Rubin 1977). The root-finding method involves inverting a square matrix of second derivatives of order $pg + \frac{1}{2}p(p+1) + g - 1$. Since the EM algorithm involves solving the likelihood equations directly, iterating from some starting value until a criterion of convergence is satisfied, it is the preferred method. Day (1969) derived an algebraic modification of the original equations that requires less computation per iteration, but unfortunately is appropriate only for two groups.

Day (1969) also mentioned the difficulty that there will be several local maxima for almost every data situation. Everitt and Hand (1980, p. 47) stated that "$\cdots$ in the multivariate situation satisfactory initial estimates are almost essential if one is to avoid misleading solutions." However, Peters and Walker (1977, p. 365) stated that "$\cdots$ with probability 1 as N approaches infinity, this procedure converges locally to the strongly consistent maximum-likelihood estimate." They defined local convergence to a limit as convergence to that limit whenever the starting values are "sufficiently near" the limit. In fact, Peters and Walker proved local convergence to the m.l.e. for a more general scheme that includes the simple scheme of using the maximum likelihood equations directly. Thus, although in small samples one must worry about finding the m.l.e. (perhaps by using the algorithm with different starting values), asymptotically one is safe. Of course, only local convergence to the m.l.e. has been proven; thus, there is still the problem of choosing "good" starting values.

Since the uncertain labels equations also fit into the framework of the EM algorithm, it makes sense to use that type of algorithm in solving them. In addition, by analogy with the theory presented for the simple mixture case, it is reasonable that this type of algorithm will have good

properties and will converge locally to the m.l.e. with probability 1 as n goes to infinity. Of course, the same difficulties concerning starting values arise.

The problem of finding the solution to the maximum likelihood equations was discussed in Section 3.2. However, there is no problem in the actual computation of the estimators. If equations (3) are solved iteratively from a starting value of $\hat{\epsilon}$ that is between 0 and 1, the estimates $\hat{\epsilon}$ in all successive iterations must also be between 0 and 1; this is because the estimated $\hat{P}(\cdot|\mathbf{x}_{ik})$ will be between 0 and 1 in every iteration when $\hat{\epsilon}$ is between 0 and 1. And if the maximum occurs at $\hat{\epsilon} = 0$ or $\hat{\epsilon} = 1$, the algorithm converges to the correct value without difficulty. What happens can be described as follows (using the case of $\hat{\epsilon} = 0$ to illustrate): if the solution to the likelihood equations would result in $\hat{\epsilon} < 0$, the algorithm keeps trying to reach that value by pushing $\hat{\epsilon}$ to smaller and smaller values. However, $\hat{\epsilon}$ is constrained to be greater than or equal to zero, so the parameter estimates change less and less. Since the criterion of convergence for the algorithm is that the estimates change less than a given amount, convergence is achieved at (or very close to) $\hat{\epsilon} = 0$.

It would be possible to check whether $\partial \ell n \, L/\partial \epsilon$ evaluated at $\epsilon = 0$ is negative before allowing the iterative process to begin, in which case no iteration is necessary and the m.l.e.'s are $\hat{\epsilon} = 0$ and the associated $\hat{\boldsymbol{\mu}}_j$'s and $\hat{\boldsymbol{\Sigma}}$. However, since the algorithm converges fairly rapidly to the solution with $\hat{\epsilon} = 0$ when $\partial \ell n \, L/\partial \epsilon$ is negative, the added computational burden of evaluating the first partial derivative appears unnecessary.

## 5. AN EMPIRICAL EXAMPLE: EMPLOYMENT STATUS OF MARRIED WOMEN

### 5.1. Economic Background of the Problem and Description of the Data

The analysis presented in this section illustrates the techniques developed in the paper using the example described in Section 1 involving the employment status of married women. The discussion here will be brief; for further details see Lichtenstein (1985).

The U.S. Bureau of Labor Statistics (BLS) defines three employment status categories: employed, unemployed, and out of the labor force. The official definitions of these categories are hierarchical in the sense that "having a job takes precedence over seeking work" (Niemi 1974, p. 333)

and seeking market work takes precedence over being active in nonmarket work. The unemployed category is defined as civilians who are not employed (in market work) in the survey week but are available for work. Thus, "a housewife who is seeking employment in the market is defined as unemployed, with her search for a job taking precedence over her nonmarket work" (Niemi 1974, p. 333).

Economic theory implies that the employment categories may not be the best way of grouping women in two specific ways. First, many women labeled as being out of the labor force have chosen the "predominant female option of a nonmarket 'occupation' ... of homemaker" (Johnson 1983, p. 61) in much the same way that other women have chosen a market occupation. Thus, these women should have some characteristics that are similar to those of some women who are officially employed.

Second, there is instability in the unemployed category because of the high mobility of women into and out of the labor force. An unemployed woman may get discouraged and drop out of the labor force because the nonmarket occupation of homemaker is an acceptable option for her. In addition, a homemaker who happened to look for a job at some point in the month prior to the survey would be considered unemployed, while another homemaker with similar characteristics who did not have time to look for a job would be considered out of the labor force.

The data of this example were collected on 5,355 married women who were part of a sample of husbands and wives taken from the Current Population Survey of 1975 (since the data are published in March of the following year, this is officially the March 1976 Current Population Survey). Therefore, conclusions reached in this analysis apply only to U.S. married women living with their husbands.

### 5.2. Model Selection

Traditional discriminant analyses provide the basis for the choice of a tentative model to be used by the m.l.e. techniques. The stepwise discriminant procedure of SPSS was used, as it provided a wide range of useful output and was flexible and easy to use. The models were judged on the basis of economic theory, misclassification percentages, plots of the data in the discriminant space, and the

canonical correlations associated with the discriminant vectors.

The abbreviations UE for unemployed, EM for employed, and NL for out of the labor force will be used for the three groups. Nine observations with unusually large values for certain variables were deleted from all analyses because they distorted the choice of model while not contributing anything important. In addition, the sample was randomly divided in half because it was so large, with one half being used to validate the results obtained for the other half.

Since there are three groups and more than one variable, the discriminant space is two-dimensional (its dimension equals min(p, g−1)), and there are two discriminant vectors. Simulation studies had revealed that problems occur in ranking the discriminant vectors by the size of their associated eigenvalues (and hence in comparing specific discriminant vectors across techniques). Thus, although denoting the vectors as first or second based on the size of their associated eigenvalues may be unclear (but convenient as a notational convention), the general interpretation of the coefficients in the individual vectors is unaffected.

The final model chosen on the basis of the preliminary analyses had a reasonable condition index of 23 (see Thisted 1988, Sec. 3.5) and no serious collinearity problems. We will discuss the vectors further in Section 5.4 when we compare the traditional and m.l.e. methods.

When the m.l.e. techniques were used for the chosen model, they produced some unexpected results. Both algorithms converged to m.l.e.'s of the group means with values of the variable "race" equal to 0 or 1, with the variance of the race variable equal to 0 (which, of course, causes $\hat{\Sigma}$ to be singular). This suggested that the race variable should not (and, in fact, could not) be in the model and demonstrated that the m.l.e. technique can be used for variable selection. Further investigation revealed that removal of the race variable from the analysis did not change the results very much.

The m.l.e. algorithm was quite expensive to run on such a complex problem and such a large data set. Thus, both half samples being used were themselves randomly split, with no substantial effect on the estimates.

### 5.3. Classification Results

The estimated label errors, probabilities of misclassification, and classification results are given in Table 5.1. The classification tables are based on classifying each observation to the group for which its posterior probability is greatest. The stability of the error parameters was assessed by validating the results on other observations, which indicated that the error parameter estimates and misclassification probabilities were remarkably stable, making it possible to interpret and evaluate the error parameters with a great deal of confidence.

Since the computed chi-square (1 degree of freedom) value of 240.56 has a p-value of less than .000001, the hypothesis that the overall error rate in the labels is zero ($\epsilon = 0$) can clearly be rejected. Hence, our suggestion that the official BLS definitions of the employment categories are inaccurate for wives with husbands present seems to be correct. Furthermore, the hypothesis that the error rate in the labels is the same for all three groups ($\epsilon_1 = \epsilon_2 = \epsilon_3$) can clearly be rejected, since the computed chi-square (2 degrees of freedom) value of 447.59 has a p-value of less than .000001.

The misclassification probabilities estimated from both the original and new observations are lowest for the group-dependent error model. The large difference between the traditional and m.l.e. misclassification probabilities is due in part to the different definitions of the posterior probabilities in the two cases (the m.l.e. values incorporate the modeled prior probabilities). However, even allowing for that difference, the group-dependent error model classifies observations much more accurately. The appropriate model for this data set seems to be the group-dependent error model; hence, most attention in the rest of this section will be focused on results for this model.

The values of the $\hat{\epsilon}_j$ are extremely informative. The economic theory presented in Section 5.1 discussed the high degree of instability in the unemployed category, thus explaining the large value of $\hat{\epsilon}_2$, the error in the UE label. The large value of $\hat{\epsilon}_3$, the error in the EM label, indicates that employed women are very heterogeneous in age, educational level, presence of young children at home, and family wages. The size of $\hat{\epsilon}_1$ implies that those women labeled NL are extremely homogeneous, as might be expected. The combination of $\hat{\epsilon}_1$ and $\hat{\epsilon}_3$ supports the economic theory that suggested a similarity between women labeled NL and some women labeled EM.

Table 5.1

M.L.E. TECHNIQUE RESULTS:
ESTIMATED LABEL ERRORS AND CLASSIFICATION RESULTS

Common Error Model

Overall error in original labor force category labels: $\hat{\epsilon} = .24664$

Test statistic for null hypothesis $\epsilon = 0$: 240.56 ($p < .000001$)

Number of iterations = 47

Total probability of misclassification estimated from original observations = .11475

Total probability of misclassification estimated from new observations = .12528

Group Dependent Error Model

Error in original label NL: $\hat{\epsilon}_1 = .07445$

Error in original label UE: $\hat{\epsilon}_2 = .67135$

Error in original label EM: $\hat{\epsilon}_3 = .78346$

Test statistic for null hypothesis $\epsilon_1 = \epsilon_2 = \epsilon_3$: 447.59 ($p < .000001$)

Number of iterations = 49

Total probability of misclassification estimated from original observations = .04236

Total probability of misclassification estimated from new observations = .04429

Traditional Model

Total probability of misclassification estimated from original observations = .30785

Total probability of misclassification estimated from new observations = .32205

The classification results for the traditional discriminant analysis provide additional support for the hypothesis that the category UE is unstable. All of the women initially labeled as UE were reassigned. Of course, unequal within-group covariance matrices might also have caused this result, since classification results based on a pooled within-group covariance matrix, which these are, would not accurately account for the spread of the data in that case. In fact, a test of the homogeneity of the covariance matrices for the three groups indicates that they are very different from each other. However, the classification results computed using individual within-group covariance matrices are very similar to those computed using a pooled covariance matrix. Thus, the suggestion that the error in the label UE is large seems valid.

### 5.4. Discriminant Vector Coefficients

Table 5.2 contains the standardized discriminant vectors obtained by the three estimation techniques. The matrix of standardized vectors, S, is obtained as S = DV, where V is the output matrix of eigenvectors that satisfy the relationship $V'\hat{\Sigma}V = I$, $D = \text{diag}(\hat{\sigma}_{11},\cdots,\hat{\sigma}_{pp})$, and the $\hat{\sigma}_{jj}^2$ are the diagonal elements of $\hat{\Sigma}$.

Looking first at the traditional discriminant vectors, it can be seen that the second vector is contributing very little, since the associated eigenvalue is very small. The second vector seems to be separating the women primarily on the basis of their full-time wage.

A comparison of the traditional vectors with the group-dependent error vectors shows how different the two sets of results are. The greatest difference is that in the group-dependent error analysis, both vectors contribute substantially to the discrimination. An examination of the first vector reveals that the general pattern of coefficients is similar to that of the traditional vector, although certain variables seem more important and others less important than for the traditional vector. The second vector seems to be discriminating primarily on the basis of the women's full-time wage; this is similar to the behavior of the second traditional vector, except that the sign of the coefficient has changed.

Table 5.2

STANDARDIZED DISCRIMINANT VECTORS

Traditional Technique

| Variable | Vector 1 | Vector 2 |
|---|---|---|
| Presence of children under 6 | .56614 | −.24426 |
| Husband's full-time wage | .58001 | .17260 |
| Husband's potential unemployment payment | −.29596 | .05664 |
| Wife's full-time wage | −.71663 | .65071 |
| Wife's potential unemployment payment | .94801 | .29699 |
| Indicator for wife under 30 | −.27361 | −.13656 |
| Wife's education level | −.37591 | −.00381 |
| Eigenvalue | .32029 | .00648 |

Common $\hat{\epsilon}$ M.l.e. Technique

| Variable | Vector 1 | Vector 2 |
|---|---|---|
| Presence of children under 6 | .69924 | −.05179 |
| Husband's full-time wage | .56477 | .23718 |
| Husband's potential unemployment payment | −.06158 | −.77810 |
| Wife's full-time wage | −.65495 | .04530 |
| Wife's potential unemployment payment | .41490 | 1.18710 |
| Indicator for wife under 30 | −.37262 | .00815 |
| Wife's education level | −.46189 | −.02168 |
| Eigenvalue | .57355 | 1.39822 |

Group-Dependent $\hat{\epsilon}_j$ M.l.e. Technique

| Variable | Vector 1 | Vector 2 |
|---|---|---|
| Presence of children under 6 | .08938 | .01443 |
| Husband's full-time wage | .35163 | −.02489 |
| Husband's potential unemployment payment | −.68733 | .35267 |
| Wife's full-time wage | −.68085 | −1.01154 |
| Wife's potential unemployment payment | 1.44293 | −.02645 |
| Indicator for wife under 30 | −.01345 | .05692 |
| Wife's education level | −.13030 | −.02574 |
| Eigenvalue | 1.08089 | 2.26460 |

### 5.5. Plots of the Discriminant Spaces

For the sake of brevity, we merely summarize these results here. Plots of the group means in the traditional and m.l.e. discriminant spaces can provide insight in several ways. First, these plots reveal that the distances among the group means are small compared to the spread of the data for all three techniques. Hence, the results must be viewed somewhat judiciously because the m.l.e. techniques do not perform as well when the means are close together as when the means are well separated relative to the within-group spread.

Second, the configuration of the means helps in the interpretation of the discriminant vectors and the characterization of the groups. The traditional means are practically on a straight line, which is in accordance with the previous conclusion that the second vector was not contributing very much; it is clear from the plot that the first vector separates the NL women from the rest.

The group-dependent error means are quite widely separated and clearly span a two-dimensional space. One of the means is very close to the mean of the initial NL group; this is in accordance with the small $\hat{\epsilon}_1$. The means of the other two groups are reversed in their spatial configuration from that of the corresponding traditional means; this is due to the change in sign of the key coefficient in vector 2 (see Table 5.2).

It is also of great interest to compare the initial label and final label of each observation by examining plots of the individual data points transformed into the different discriminant spaces. The plot of the data transformed into the group-dependent error m.l.e. space indicates that the three groups corresponding to the three means are quite distinct. It seems that the women originally labeled EM have been split into two groups, while the women originally labeled NL are still grouped together and concentrated in approximately the same location as in the traditional space.

### 5.6. Conclusions

Based on the analysis presented here, it seems that the official employment categories do not reflect the true employment status groupings of married women. It must be mentioned that there is another possible explanation for the results presented in this section. If the population within-group covariance matrices are not equal, the unequal variation of the groups can be confused with the label

errors. In that case, the estimated common within-group covariance matrix, $\hat{\Sigma}$, is really estimating an "average" of the three population within-group covariance matrices. Observations that are not far from the mean of the group defined by their uncertain labels in the metric of the correct covariance matrix for that population may appear far in the metric of the estimated average of the covariance matrices, thus confusing the m.l.e. algorithm into considering them mislabeled (which will, of course, result in an incorrectly inflated estimate of $\epsilon_j$).

The sample within-group covariance matrices are different for the three groups defined by the original uncertain labels and the hypothesis that the population within-group covariance matrices are the same is rejected based on these sample covariance matrices. However, at the present time there is no way to determine whether the population within-group covariance matrices are equal if the labels truly are uncertain. Either the original labels are correct and the inequality of the population covariance matrices is causing the appearance of label errors, or the original labels are uncertain and the equality of the population covariance matrices is still undetermined. In the latter case, unequal population covariance matrices might still affect the estimated label errors, in directions depending on the relative sizes of the three within-group covariance matrices.

Assuming that the population covariance matrices are all the same, the analyses presented in this section demonstrate how the techniques proposed in this paper can improve on the traditional discriminant techniques. Valid group-dependent error rates in the labels were estimated with associated discriminant vectors that produced lower misclassification probabilities.

Our results agree with the conclusion reached by Johnson (1983) that the definitions and methodology used to determine unemployment status categories have considerable influence on unemployment statistics for women. In the analysis described in this section, many of the women labeled EM had characteristics that grouped them together with women labeled NL. One can interpret this as supporting Johnson's view that home production can be recognized as one form of employment. Thus, the m.l.e. techniques proposed in this paper uncover a phenomenon suspected in other studies and help to redefine groupings hypothesized to be inaccurately defined.

## 6. CONCLUSION

The problem discussed in this paper, uncertain group assignments in a discriminant setting, is much more widespread than is generally recognized. Analysts usually ignore the uncertainty in the groupings and utilize existing statistical techniques. There is a need for a method that explicitly accounts for the error in the group labels and estimates its prevalence.

The statistical technique suggested here can be recommended for several reasons. First, the model of label uncertainty is intuitively appealing, and the parameter estimators are intuitively sensible. Second, the parameter estimators have good asymptotic properties, which the traditional estimators, computed by ignoring the uncertainty in the labels, do not. Finally, the computer algorithm proposed is reasonably straightforward, since it is a form of EM algorithm, and appears to have good properties. For the first time, problems involving uncertainly grouped observations can be analyzed properly. The properties of the techniques discussed in this paper have been shown to support the worth and practicality of these techniques.

## References

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis* (2nd ed.), New York: John Wiley & Sons.

Ashikaga, T., and Chang, P. C. (1981), "Robustness of Fisher's Linear Discriminant Function Under Two-Component Mixed Normal Models," *Journal of the American Statistical Association*, 76, 676-680.

Bradley, R. A., and Gart, J. J. (1962), "The Asymptotic Properties of ML Estimators When Sampling From Associated Populations," *Biometrika*, 49, 205-214.

Chittineni, C. B. (1982), "Maximum Likelihood Estimation of Label Imperfection Probabilities and Its Use in the Identification of Mislabeled Patterns," *IEEE Transactions on Geoscience and Remote Sensing*, GE20, 99-111.

Day, N. E. (1969), "Estimating the Components of a Mixture of Normal Distributions," *Biometrika*, 56, 463-474.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

Everitt, B. S., and Hand, D. J. (1980), *Finite Mixture Distributions*, New York: Chapman and Hall.

Ganesalingam, S., and McLachlan, G. J. (1978), "The Efficiency of a Linear Discriminant Function Based on Unclassified Initial Samples," *Biometrika*, 65, 658-662.

Ganesalingam, S., and McLachlan, G. J. (1979), "Small Sample Results for a Linear Discriminant Function Estimated From a Mixture of Normal Populations," *Journal of Statistical Computation and Simulation*, 9, 151-158.

Ganesalingam, S., and McLachlan, G. J. (1980), "Error Rate Estimation on the Basis of Posterior Probabilities," *Pattern Recognition*, 12, 405-413.

Geisser, S. (1982), "Bayesian Discrimination," in *Handbook of Statistics* (Vol. 2), P. R. Krishnaiah and L. N. Kanal (Eds.), Amsterdam: North-Holland, pp. 101-120.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, New York: John Wiley & Sons.

Hand, D. J. (1981), *Discrimination and Classification*, New York: John Wiley & Sons.

Hassleblad, V. (1966), "Estimation of Parameters for a Mixture of Normal Distributions," *Technometrics*, 8, 431-446.

Hosmer, D. W. (1973), "On Maximum Likelihood Estimation of the Parameters of a Mixture of Two Normal Distributions When the Sample Size is Small," *Communications in Statistics*, 1, 217-227.

Johnson, J. (1983), "Sex Differentials in Unemployment Rates: a Case for No Concern," *Journal of Political Economy*, 91, 293-303.

Kendall, M. G., and Stuart, A. (1979), *The Advanced Theory of Statistics* (Vol. 2, 4th ed.), New York: Macmillan Publishing Co.

Kshirsagar, A. M. (1972), *Multivariate Analysis*, New York: Marcel Dekker, Inc.

Lachenbruch, P. A. (1966), "Discriminant Analysis When Initial Samples Are Misclassified," *Technometrics*, 8, 657-662.

Lachenbruch, P. A. (1974), "Discriminant Analysis When Initial Samples Are Misclassified: II. Nonrandom Misclassification Models," *Technometrics*, 16, 419-424.

Lachenbruch, P. A. (1975), *Discriminant Analysis*, New York: Hafner Press.

Lichtenstein, C. H. (1985), "Discriminant Analysis When the Group Labels Are Uncertain," unpublished Ph.D. dissertation, Cornell University, Ithaca, NY.

McLachlan, G. J. (1972), "Asymptotic Results for Discriminant Analysis When the Initial Samples Are Misclassified," *Technometrics*, 14, 415-422.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: John Wiley & Sons.

Niemi, B. (1974), "The Female-Male Differential in Unemployment Rates," *Industrial and Labor Relations Review*, 27, 331-350.

Peters, B. C., and Walker, H. F. (1978), "An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions," *SIAM Journal of Applied Mathematics*, 35, 362-378.

Shanmugam, K., and Breiphol, A. M. (1971), "An Error Correcting Procedure for Learning With an Imperfect Teacher," *IEEE Transactions on Systems, Man and Cybernetics*, 1, 223-229.

Thisted, R. A. (1988), *Elements of Statistical Computing*, New York: Chapman and Hall.

Tyler, D. E. (1981), "Asymptotic Inference for Eigenvectors," *Annals of Statistics*, 9, 725-736.

Wald, A. (1949), "Note on the Consistency of the Maximum Likelihood Estimate," *Annals of Mathematical Statistics*, 20, 595-601.

## APPENDIX: PROOFS

*Proof of Theorem 1.*

The strong consistency of the m.l.e.'s is proved by an extension of the proof in Wald (1949). The following assumptions in addition to those in Wald are necessary:

1. The parameters are identifiable.
2. $\Sigma$ is nonsingular.
3. There exist constants $\tilde{q}_1,\cdots,\tilde{q}_g$ such that $\tilde{n}_k/n \to \tilde{q}_k$ as $n_k \to \infty$.

Wald's proof is for the case of i.i.d. observations; we extend his proof to the case of g groups of i.i.d. observations. Assumption 1 above is equivalent to Assumption 4 of Wald. Assumptions 1–8 in Wald's paper hold for each of the $f_k(\mathbf{x})$ if his Assumptions 3, 5, and 7 are replaced by Assumption 9 given at the end of that paper. Lemmas 1–3 of Wald's paper hold for each $f_k(\mathbf{x})$; this follows directly from the assumptions of these lemmas.

Wald proved the result in two steps. Rewrite his Theorem 1 (the first step) as

$$P\left\{ \lim_{\substack{n\to\infty \\ \tilde{n}_k/n\to\tilde{q}_k}} \frac{\sup_{\boldsymbol{\theta}\in\omega} \prod_{k=1}^{g} \prod_{i=1}^{\tilde{n}_k} f_k(\mathbf{x}_{ik};\boldsymbol{\theta})}{\prod_{k=1}^{g} \prod_{i=1}^{\tilde{n}_k} f_k(\mathbf{x}_{ik};\boldsymbol{\theta}^\circ)} = 0 \right\} = 1 ,$$

where $\omega$ is any closed subset of the parameter space that does not contain the true value $\boldsymbol{\theta}^\circ$. Then the proof follows directly from the strong law of large numbers and Slutsky's Theorem, just as in Wald's paper. Wald's Theorem 2 (the second step of the proof) applies with no changes, so the strong consistency of the m.l.e.'s is proved.

*Proof of Theorem 2.*

The proof of the asymptotic normality of the m.l.e.'s requires the following additional assumptions:

4. For almost all $\mathbf{x}_k \in R_k$ and for all $\boldsymbol{\theta} \in \Theta$,

$$\frac{\partial \ell n\, f_k}{\partial \theta_r}, \quad \frac{\partial^2 \ell n\, f_k}{\partial \theta_r \partial \theta_s}, \quad \text{and} \quad \frac{\partial^3 \ell n\, f_k}{\partial \theta_r \partial \theta_s \partial \theta_t}$$

exist for r, s, t = 1,$\cdots$, $pg + \frac{1}{2}p(p+1) + 1$; k = 1,$\cdots$,g.

5. For almost all $\mathbf{x}_k \in R_k$ and for every $\boldsymbol{\theta} \in \Theta$,

$$\left|\frac{\partial f_k}{\partial \theta}\right| < F_{kr}(\mathbf{x}_k), \quad \left|\frac{\partial^2 f_k}{\partial \theta_r \partial \theta_s}\right| < F_{krs}(\mathbf{x}_k)$$

and

$$\frac{\partial^3 \ell n\, f_k}{\partial \theta_r \partial \theta_s \partial \theta_t} < H_{krst}(\mathbf{x}_k)$$

where $F_{kr}(\mathbf{x}_k)$ and $F_{krs}(\mathbf{x}_k)$ are integrable over $R_k$, and

$$\int_{R_k} H_{krst}(\mathbf{x}_k)\, f_k d\mathbf{x}_k < M_k$$

($k = 1,\cdots,g$; $r, s, t = 1,\cdots,pg + \frac{1}{2}p(p+1) + 1$), where the $M_k$ are finite positive constants.

6. For all $\boldsymbol{\theta} \in \Theta$, the matrix $J = [J_{rs}(\boldsymbol{\theta})]$ is positive definite with finite determinant.

The proof of part (i) follows very closely the proof of Theorem 1(iv) of Bradley and Gart (1962), which develops the asymptotic properties of mixture distributions. Letting $\boldsymbol{\theta} = (\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_g, \Sigma, \epsilon)$, this theorem states that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ) \rightarrow \mathbf{Z}$ in distribution as $n \rightarrow \infty$ and $\tilde{n}_k/n \rightarrow \tilde{q}_k$, where $\mathbf{Z} \sim \text{MVN}(\mathbf{0}, \mathbf{J}_0^{-1})$ and $\boldsymbol{\theta}^\circ$ is the vector of true parameter values. The proof is based on expansions of the partial derivatives of the log likelihood function $\ell n\, L = \sum_{k=1}^{g} \sum_{i=1}^{\tilde{n}_k} \ell n\, f_k(\mathbf{x}_{ik})$.

The proof of part (ii) follows the proof of part (i) with one modification. As before, the quantities $[\partial \ell n\, f_k(\mathbf{x}_{ik})/\partial \theta_r]_{\epsilon = 0}$, $i = 1,\cdots,\tilde{n}_k$, are i.i.d. values that can be negative. Hence

$$\sum_{k=1}^{g} \sqrt{\frac{\tilde{n}_k}{n}} \sum_{i=1}^{\tilde{n}_k} \frac{1}{\sqrt{\tilde{n}_k}} \left( \frac{\partial \ell n\, f_k(\mathbf{x}_{ik})}{\partial \theta_r} \right)_{\epsilon = 0}$$

has an asymptotic marginal normal distribution. However, $\epsilon$ is restricted to be greater than or equal to zero, so $\hat{\epsilon}$ itself cannot have an asymptotic marginal normal distribution, and the joint asymptotic distribution of all the estimators is restricted to being nonnegative in the $\hat{\epsilon}$ dimension. Note that although the matrix $\mathbf{J}_0^*$ is equal to the upper left square submatrix of $\mathbf{J}_0$ of dimension $pg + \frac{1}{2}p(p+1)$, $(\mathbf{J}_0^*)^{-1}$ and the corresponding upper left square submatrix of $\mathbf{J}_0^{-1}$ are not necessarily the same, because $\epsilon$ has not been shown to be asymptotically independent of the other parameters.

*Proof of Corollary*

Since $\hat{\Delta}_{\hat{q}} \rightarrow \Delta_q$, basic results about limits ensure that $\hat{\Sigma}^{-1}\hat{\Delta}_{\hat{q}} \rightarrow \Sigma^{-1}\Delta_q$ in probability. Since $\hat{\Sigma}\,\hat{\Sigma}^{-1}\hat{\Delta}_{\hat{q}} = \hat{\Delta}_{\hat{q}}$ is symmetric and $\hat{\Sigma}$ is positive definite symmetric, Lemma 2.1 of Tyler (1981) ensures that $\hat{\lambda}_j \rightarrow \lambda_j$ in probability, $j = 1,\cdots,r$, and that the total eigenprojection associated with $\hat{\lambda}_1,\cdots,\hat{\lambda}_r = \hat{p} \rightarrow P$ = total eigenprojection associated with $\lambda_1,\cdots,\lambda_r$ in probability, as long as $\lambda_1,\cdots,\lambda_r$ are distinct.

*Proof of Theorem 3*

The proof of the asymptotic distribution of $-2\ \ell n\ \lambda$ for the i.i.d. case, where $\lambda$ is the generalized likelihood ratio statistic, can be found in Kendall and Stuart (1979, pp. 246-247). This proof, which assumes the asymptotic multivariate normality of the estimators, applies directly to the uncertain labels case. The only modification necessary, because of the truncated distribution of $\hat{\epsilon}$ when $\epsilon = 0$, is that the critical value used is $\chi^2_1(1-2\alpha)$ instead of the value $\chi^2_1(1-\alpha)$ given in Kendall and Stuart. This change is required by the fact that the hypothesis will never be rejected when $\hat{\epsilon} = 0$.

*Proof of Theorem 4*

Since $\mathbf{x}_{1k},\cdots,\mathbf{x}_{\tilde{n}_k,k}$ are i.i.d., $(1/\tilde{n}_k)\ \sum_{i=1}^{\tilde{n}_k} \mathbf{x}_{ik}\mathbf{x}'_{ik} \rightarrow E(\mathbf{x}_k\mathbf{x}'_k)$ with probability 1 and $\bar{\mathbf{x}}_k \rightarrow E(\mathbf{x}_k)$ with probability 1 as $\tilde{n}_k \rightarrow \infty$, by the strong law of large numbers. Thus, it can be shown using Slutsky's Theorem and certain algebraic identities that

$$\frac{n-g}{n}\,\tilde{\mathbf{W}} \rightarrow \sum_{k=1}^{g} \tilde{q}_k\Big[\Sigma + (1-\epsilon)\mu_k\mu'_k + \frac{\epsilon}{g-1}\sum_{\ell\neq k}\mu_\ell\mu'_\ell\Big] - \sum_{k=1}^{g}\tilde{q}_k E(\mathbf{x}_k)E(\mathbf{x}_k)'$$

$$= \Sigma + \sum_{j=1}^{g} q_j\mu_j\mu'_j - \sum_{k=1}^{g}\tilde{q}_k E(\mathbf{x}_k)E(\mathbf{x}_k)' = \Sigma + \tilde{\mathbf{U}}_\epsilon$$

with probability 1 as $\tilde{n}_k \rightarrow \infty$ and $\tilde{n}_k/n \rightarrow \tilde{q}_k$. Similarly, it can be shown that

$$\frac{g-1}{n}\,\tilde{\mathbf{B}} \rightarrow \sum_{k=1}^{g}\tilde{q}_k E(\mathbf{x}_k)E(\mathbf{x}_k)'$$

$$-\Big[\sum_{k=1}^{g}\tilde{q}_k\Big((1-\epsilon)\mu_k + \frac{\epsilon}{g-1}\sum_{\ell\neq k}\mu_\ell\Big)\Big]\Big[\sum_{k=1}^{g}\tilde{q}_k\Big((1-\epsilon)\mu_k + \frac{\epsilon}{g-1}\sum_{\ell\neq k}\mu_\ell\Big)\Big]'$$

$$= \Delta_q - \tilde{\mathbf{U}}_\epsilon$$

with probability 1 as $\tilde{n}_k \rightarrow \infty$ and $\tilde{n}_k/n \rightarrow \tilde{q}_k$. Now, $\tilde{\mathbf{U}}_\epsilon$ can be written as

$$\tilde{\mathbf{U}}_\epsilon = \sum_{k=1}^{g}\tilde{q}_k\epsilon(1-\epsilon)(\mu_k-\bar{\mu}_{(k)})(\mu_k-\bar{\mu}_{(k)})'$$

$$+ \sum_{k=1}^{g}\tilde{q}_k\frac{\epsilon}{g-1}\sum_{\ell\neq k}(\mu_\ell-\bar{\mu}_{(k)})(\mu_\ell-\bar{\mu}_{(k)})'$$

where $\bar{\boldsymbol{\mu}}_{(k)} = [1/(g-1)] \sum_{\ell \neq k} \boldsymbol{\mu}_\ell$. Hence $\tilde{\mathbf{U}}_\epsilon$ is nonnegative definite since it can be written as the sum of nonnegative definite matrices. Of course, $\tilde{\mathbf{U}}_\epsilon$ should be nonnegative definite to ensure that $\boldsymbol{\Sigma} + \tilde{\mathbf{U}}_\epsilon$ is a covariance matrix.

*Proof of Theorem 5*

The estimated error rate of (6) can be rewritten as

$$\hat{E} = 1 - \sum_{k=1}^{g} \frac{\tilde{n}_k}{n} \cdot \frac{1}{\tilde{n}_k} \sum_{i=1}^{\tilde{n}_k} \hat{P}(j|\mathbf{x}_{ik})$$

where $\hat{P}(j|\mathbf{x}_{ik}) = \max_m \hat{P}(m|\mathbf{x}_{ik})$. By the weak law of large numbers, Slutsky's Theorem, and the consistency of the $\hat{P}(\cdot|\mathbf{x})$,

$$\hat{E} \rightarrow 1 - \sum_{k=1}^{g} \tilde{q}_k E_{\mathbf{x}_k}[\, P(j|\mathbf{x}_k)] = \sum_{k=1}^{g} \tilde{q}_k (1 - E_{\mathbf{x}_k}[\, P(j|\mathbf{x}_k)]) \quad = \tilde{R}$$

in probability as $n \rightarrow \infty$ and $\tilde{n}_k/n \rightarrow \tilde{q}_k$.

*Proof of Theorem 6*

The proof of part (i) follows that of Theorem 3. The proof of part (ii) follows that of Theorem 3 also, except that a $1-\alpha$ critical value is used because the hypothesis no longer involves the boundary of the parameter space. In addition, since there are now g–1 restrictions on the parameters, the distribution has g–1 degrees of freedom instead of only one (Kendall and Stuart 1979, pp. 246-247).