Lockheed Electronics Company, Inc.
A SUBSIDIARY OF LOCKHEED CORPORATION
1830 NASA Road 1, Houston, Texas 77058
Tel. 713-333-5411

JSC-16067
SEP 1979
NASA CR-160364

Ref: 642-7734
Contract NAS 9-15800
Job Order 73-705

TECHNICAL MEMORANDUM

MAXIMUM LIKELIHOOD ESTIMATION OF LABEL IMPERFECTIONS AND ITS USE IN THE IDENTIFICATION OF MISLABELED PATTERNS

By

C. B. Chittineni

(NASA-CR-160364) MAXIMUM LIKELIHOOD ESTIMATION OF LABEL IMPERFECTIONS AND ITS USE IN THE IDENTIFICATION OF MISLABELED PATTERNS (Lockheed Electronics Co.) 44 p HC A03/MF A01. N80-12805
Unclas
CSCL 12A G3/65 40331

Approved By: J. C. Minter
T. C. Minter, Supervisor
Techniques Development Section

~~September 1979~~ LEC-13678

DISTRIBUTION

Distribution of this document is limited to those people whose names are followed by an asterisk in the following list; all others receive an abstract (JSC Form 1424) only.†

JSC/G. Badhwar/SF3*
K. Baker/SF3*
R. R. Baldwin/SF3
T. L. Barnett/SF3
R. M. Bizzell/SF4*
I. D. Browne/SF3
L. F. Childs/SF2
K. J. Demel/SF3
H. G. deVezin/FM8
J. W. Dietrich/SF3
J. L. Dragg/SF4*
R. B. Erb/SF1*
J. D. Erickson/SF3*
J. G. Garcia/SF3
G. E. Graybeal/SF5*
F. G. Hall/SF1*
C. R. Hallum/SF4*
K. E. Henderson/SF3
W. E. Hensley/SF2
R. P. Heydorn/SF3*
R. O. Hill/SF4
A. G. Houston/SF4*
R. D. Juday/SF3
T. W. Pendleton/SF3*
D. E. Pitts/SF3*
R. G. Stuff/SF3
D. R. Thompson/SF3
M. C. Trichel/SF3*
V. S. Whitehead/SF3

USDA/G. O. Boatwright
A. D. Frank/SF6
R. E. Hatch/SA4
J. D. Murphy
R. L. Packard/SA4*

NOAA/J. D. Hill/SF2
D. G. McCrary/SF4

LEC/J. G. Baron
M. L. Bertrand
B. L. Carroll*
J. E. Davis
P. L. Krumm
P. C. Swanzy
J. J. Vaccaro
Data Research and Control (3)*
Technical Library (5)*
Job Order File*

ERIM/Q. A. Holmes*
R. Horvath
D. Rice

KSU/A. M. Feyerherm
E. T. Kanemasu

LARS/M. E. Bauer*
D. A. Landgrebe*
T. L. Phillips
P. H. Swain*

TAMU/L. F. Guseman*
J. C. Harlan
H. O. Hartley

UCB/R. N. Colwell
C. M. Hay*
R. W. Thomas

UH/H. P. Decell*

ESCS/W. H. Wigton

---

†To obtain a copy of this document, contact one of the following:

J. D. Erickson — NASA/JSC Supporting Research Branch (SF3)

J. E. Wainwright — LEC/SSD EO Development and Evaluation Department (626-42, C09)

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
JSC-16067		

4. Title and Subtitle	5. Report Date
Maximum Likelihood Estimation of Label Imperfections and Its Use in the Identification of Mislabeled Patterns	August 1979
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
C. B. Chittineni Lockheed Electronics Company, Inc.	LEC-13678
	10. Work Unit No.

9. Performing Organization Name and Address	11. Contract or Grant No.
Lockheed Electronics Company, Inc. Systems and Services Division 1830 NASA Road 1 Houston, Texas 77058	NAS 9-15800
	13. Type of Report and Period Covered

12. Sponsoring Agency Name and Address	14. Sponsoring Agency Code
National Aeronautics and Space Administration Lyndon B. Johnson Space Center Houston, Texas 77058 Technical Monitor: J. D. Erickson/SF3	Technical Memorandum

15. Supplementary Notes

16. Abstract

This paper presents the problem of estimating label imperfections and the use of the estimation in identifying mislabeled patterns. Expressions for the maximum likelihood estimates of classification errors and a priori probabilities are derived from the classification of a set of labeled and unlabeled patterns. Expressions also are presented for the asymptotic variances of probability of correct classification and proportions. Simple models are developed for imperfections in the labels and for classification errors and are used in the formulation of a maximum likelihood estimation scheme. Schemes are presented for the identification of mislabeled patterns in terms of thresholds on the discriminant functions for both two-class and multiclass cases. Expressions are derived for the probability that the imperfect label identification scheme will result in a wrong decision and are used in computing thresholds. Furthermore, the results of practical applications of these techniques in the processing of remotely sensed multispectral data are presented.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Asymptotic variance, classification errors, discriminant functions, incorrect label identification, labeling errors, linear classifier, maximum likelihood estimates, optimization, percentage of correct classification, proportions, thresholds.	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	45	

CONTENTS

TABLES

FIGURES

1. INTRODUCTION

In the practical applications of pattern recognition (such as in the processing of remotely sensed imagery data), obtaining labels is a difficult problem. Acquiring labels is expensive, and very often these labels are imperfect.

Several scientists have investigated the problem of pattern recognition with imperfectly labeled patterns (refs. 1-7). Duda and Singleton (ref. 1) showed that, for orthogonal pattern vectors, the average weight vector of a threshold logic unit converges to a solution weight vector for the correctly labeled pattern set. Kashyap (ref. 2) proposed an iterative training procedure for a two-class case. Shanmugam and Breiphol (ref. 3) developed an error-correcting procedure for disjoint densities using Parzen estimators. Chittineni (refs. 4-7) investigated the problem of learning with imperfectly labeled patterns and studied the applicability of probabilistic distance measures for feature selection with imperfectly labeled patterns. Most of these proposed schemes require the knowledge of probabilities of label imperfections, which usually are not available.

Several authors considered the problem of estimating recognition system performance (refs. 8-13). Highleyman (ref. 8) investigated the problem of estimating the probability of error of a given classifier both for known and unknown a priori probabilities. Fukunaga and Kessell (ref. 9) examined the problem of estimating the probability error from unclassified samples. Havens et al. (ref. 10) reported the experimental results of estimating the probability of error from unclassified samples using remotely sensed agricultural data. Chow (ref. 11) established a relationship between error and rejection rates which is useful in estimating the probability of error from unclassified samples.

In practice, the situation often arises in which a set of imperfectly labeled test patterns and a set of unlabeled patterns are available. (For example, in remote sensing, a set of labeled patterns called type 2 dots and a set of unlabeled patterns are usually available). This paper presents the problem of

estimating recognition system performance and label imperfections as maximum likelihood estimates from the classifier decisions of labeled and unlabeled patterns. The probabilities of the estimated label imperfections are then used in developing schemes for the identification of mislabeled patterns. The paper is organized in the following manner.

Assuming no imperfections in the labels, expressions are derived for the maximum likelihood estimates of probability of error, probability of correct classification, and a priori probabilities (section 2); also, in this section, expressions are derived for the asymptotic variances of probability of correct classification and a priori probabilities. In section 3, imperfections in the labels are introduced, models for the label imperfections and probabilities of errors are developed, and the simulation results from the processing of remotely sensed data are presented. Methods of identifying mislabeled patterns for both two-class and multiclass cases are reported in section 4, and the results of their applications in processing remotely sensed data are described. Conclusions are presented in section 5.

2. MAXIMUM LIKELIHOOD ESTIMATION OF PROBABILITY OF ERROR, PROBABILITY OF CORRECT CLASSIFICATION, AND A PRIORI PROBABILITIES

In this section, expressions are derived for the maximum likelihood estimates of probability of error, probability of correct classification, and proportions. Also, expressions for the asymptotic variance of probability of correct classification and proportion estimates are derived. It is assumed that the classifier is designed and the classifier classifications of a set of labeled and unlabeled patterns are obtained. [In a situation involving remote sensing, the labeled patterns are the test set or type 2 dots and the unlabeled patterns are the spectral values of the picture elements (pixels) for which no labels are available.] In this section, the labels of the test patterns are assumed perfect; in section 3, the labels are assumed to be imperfect. The classifier classifications of the labeled and unlabeled sets are illustrated in table 2-1.

Let ω be the given label and ω_c be the classifier label. Let $\lambda_{ij} = P(\omega = i|\omega_c = j)$ be the probability that the true label is i, given that the classifier label is j. Let $p_{ij} = P(\omega = i,\omega_c = j)$ be the probability that the true label of the pattern is i and the classifier label is j. Let $P_c(i) = P(\omega_c = i)$ be the probability that the classifier classifies a pattern into class i and $P_i = P(\omega = i)$ be the a priori probability of class i. Then we obtain

$$\begin{aligned} p_{ij} &= P(\omega = i,\omega_c = j) \\ &= P(\omega_c = j)P(\omega = i|\omega_c = j) \\ &= P_c(j)\lambda_{ij} \end{aligned} \tag{2-1}$$

TABLE 2-1.— CLASSIFICATIONS OF LABELED AND UNLABELED SETS

(a) Confusion matrix of labeled test set

True label	Classifier label				Number belonging to each class
	1	2	•••	M	
1	m_{11}	m_{12}	•••	m_{1M}	$m_{1.}$
2	m_{21}	m_{22}	•••	m_{2M}	$m_{2.}$
⋮	⋮	⋮		⋮	⋮
M	m_{M1}	m_{M2}	•••	m_{MM}	$m_{M.}$
Number classified into each class	$m_{.1}$	$m_{.2}$	•••	$m_{.M}$	$m = m_{..}$

(b) Matrix of classifications of unlabeled set

Classifier label			
1	2	•••	M
X_1	X_2	•••	X_M

where

m_{ij} = number of labeled patterns for which the true or given label is i and the classifier label is j

M = number of classes

$m_{i.} = \sum_{j=1}^{M} m_{ij}$

$m_{.j} = \sum_{i=1}^{M} m_{ij}$

$m = m_{..} = \sum_{i=1}^{M} \sum_{j=1}^{M} m_{ij}$, the total number of labeled patterns

X_j = number of unlabeled patterns for which the classifier label is j

Since each classification is independent, the likelihood function of the observed m's and X's can be written as

$$L = C \prod_{i=1}^{M} \prod_{j=1}^{M} (p_{ij})^{m_{ij}} \prod_{j=1}^{M} [P_c(j)]^{X_j}$$

$$= C \prod_{i=1}^{M} \prod_{j=1}^{M} (\lambda_{ij})^{m_{ij}} \prod_{j=1}^{M} [P_c(j)]^{X_j + m_{.j}} \qquad (2\text{-}2)$$

where C is a constant. The constraints on λ_{ij} and $P_c(j)$ are

$$\left.\begin{aligned} \sum_{i=1}^{M} \lambda_{ij} &= 1 \quad ; \quad j = 1,2,\cdots,M \\ \sum_{j=1}^{M} P_c(j) &= 1 \end{aligned}\right\} \qquad (2\text{-}3)$$

The objective is to find the values for λ_{ij} and $P_c(j)$ which maximize L, subject to the constraints of equation (2-3). Since the logarithm is a monotonic function of its argument, taking the logarithm of L and introducing Lagrangian multipliers yields

$$L' = \log C + \sum_{i=1}^{M} \sum_{j=1}^{M} m_{ij} \log(\lambda_{ij}) + \sum_{j=1}^{M} (X_j + m_{.j}) \log[P_c(j)]$$

$$+ \sum_{j=1}^{M} r_j \left(\sum_{i=1}^{M} \lambda_{ij} - 1 \right) + s \left[\sum_{j=1}^{M} P_c(j) - 1 \right] \qquad (2\text{-}4)$$

where r_j $(j = 1,2,\cdots,M)$ and s are Lagrangian multipliers. Differentiating L' with respect to $P_c(j)$ and s, equating the resulting expressions to zero, and solving for $P_c(j)$ results in

$$\hat{P}_c(j) = \frac{m_{.j} + X_j}{\sum_{\ell=1}^{M} (m_{.\ell} + X_\ell)} \qquad (2\text{-}5)$$

Similarly, the maximum likelihood estimate of λ_{ij} can be obtained as

$$\hat{\lambda}_{ij} = \frac{m_{ij}}{m_{.j}} \tag{2-6}$$

From the invariance property of the maximum likelihood estimators, the maximum likelihood estimate $\hat{P}_{cc}$ for the probability of correct classification P_{cc} can be obtained from the expression

$$\begin{aligned} P_{cc} &= \sum_{i=1}^{M} P(\omega = i, \omega_c = i) \\ &= \sum_{i=1}^{M} P(\omega_c = i) P(\omega = i | \omega_c = i) \\ &= \sum_{i=1}^{M} P_c(i) \lambda_{ii} \end{aligned} \tag{2-7}$$

Using equations (2-5) and (2-6) in equation (2-7) yields

$$\hat{P}_{cc} = \frac{\sum_{i=1}^{M} \frac{m_{ii}}{m_{.i}} (m_{.i} + X_i)}{\sum_{\ell=1}^{M} (m_{.\ell} + X_\ell)} \tag{2-8}$$

An intuitive justification for $\hat{P}_{cc}$ may be given as follows. The ratio $(m_{ii}|m_{.i})$ gives the proportion of the patterns truly belonging to class i to the patterns classified into class i. Multiplying this ratio by $(m_{.i} + X_i)$ and summing it from 1 to M gives an estimate for the number of correctly classified patterns from all patterns in the classified classes. The estimate of P_{cc} is then divided by the total number of patterns. An estimate $\hat{P}_i$ for the proportion P_i may be obtained as follows.

$$P_i = P(\omega = i)$$

$$= \sum_{j=1}^{M} P(\omega = i, \omega_c = j)$$

$$= \sum_{j=1}^{M} P(\omega_c = j) P(\omega = i | \omega_c = j)$$

$$= \sum_{j=1}^{M} P_c(j) \lambda_{ij} \qquad (2\text{-}9)$$

From equations (2-5), (2-6), and (2-9), the following is obtained.

$$\hat{P}_i = \frac{\sum_{j=1}^{M} \left[\frac{m_{ij}}{m_{.j}} (m_{.j} + X_j) \right]}{\sum_{\ell=1}^{M} (m_{.\ell} + X_\ell)} \qquad (2\text{-}10)$$

Different probabilities of error can be written as

$$P(\omega_c = j | \omega = i) = \frac{P(\omega_c = j) P(\omega = i | \omega_c = j)}{P(\omega = i)} \qquad (2\text{-}11)$$

Using equations (2-5), (2-6), and (2-10) in equation (2-11) obtains the maximum likelihood estimates $[\hat{P}(\omega_c = j | \omega = i)]$ for different probabilities of error.

$$\hat{P}(\omega_c = j | \omega = i) = \frac{\frac{m_{ij}}{m_{.j}} (m_{.j} + X_j)}{\sum_{\ell=1}^{M} \frac{m_{i\ell}}{m_{.\ell}} (m_{.\ell} + X_\ell)} \quad ; \quad i,j = 1,2,\cdots,M \qquad (2\text{-}12)$$

The estimate of equation (2-12) can be interpreted as follows. It is the ratio of the number of patterns that truly belong to class i but were classified into class j to the total number of patterns that truly belong to class i from the patterns classified into all classes.

In the following example, expressions are derived for the asymptotic variance of the estimates of the probability of correct classification and proportions. From equation (2-7), the estimated $\hat{P}_{cc}$ can be written as

$$\hat{P}_{cc} = \sum_{i=1}^{M} \hat{P}_c(i)\hat{\lambda}_{ii} \tag{2-13}$$

The delta method (ref. 14) is used to compute the asymptotic variance of $\hat{P}_{cc}$. This involves expanding $\hat{P}_{cc}$ in a Taylor series around the true value $P_{cc} = \sum_{i=1}^{M} P_c(i)\lambda_{ii}$. The result of this expansion is

$$\begin{aligned}
\mathrm{Var}\left(\hat{P}_{cc}\right) = & \sum_{i=1}^{M}\sum_{j=1}^{M} \mathrm{Cov}\left(\hat{\lambda}_{ii}\hat{\lambda}_{jj}\right) \frac{\partial P_{cc}}{\partial \lambda_{ii}} \frac{\partial P_{cc}}{\partial \lambda_{jj}} \\
& + \sum_{i=1}^{M}\sum_{j=1}^{M} \mathrm{Cov}\left[\hat{\lambda}_{ii}\hat{P}_c(j)\right] \frac{\partial P_{cc}}{\partial \lambda_{ii}} \frac{\partial P_{cc}}{\partial P_c(j)} \\
& + \sum_{i=1}^{M}\sum_{j=1}^{M} \mathrm{Cov}\left[\hat{P}_c(i)\hat{\lambda}_{jj}\right] \frac{\partial P_{cc}}{\partial P_c(i)} \frac{\partial P_{cc}}{\partial \lambda_{jj}} \\
& + \sum_{i=1}^{M}\sum_{j=1}^{M} \mathrm{Cov}\left[\hat{P}_c(i)\hat{P}_c(j)\right] \frac{\partial P_{cc}}{\partial P_c(i)} \frac{\partial P_{cc}}{\partial P_c(j)}
\end{aligned} \tag{2-14}$$

The number of independent parameters is 2M - 1; namely, $\lambda_{11},\lambda_{22},\cdots,\lambda_{MM}$ and $P_c(1),P_c(2),\cdots,P_c(M-1)$. If these parameters are labeled by δ_i, $i = 1,2,\cdots,2M-1$, the (2M - 1) by (2M - 1) information matrix, the general term of which is given by $E\left(-\frac{\partial^2 \log L}{\partial\delta_i \partial\delta_j}\right)$, can be evaluated from equation (2-2). Carrying out these calculations and inverting the resulting matrix yields the variance-covariance matrix of λ_{ii}, $i = 1,2,\cdots,M$, and $P_c(j)$, $j = 1,2,\cdots,M-1$. From this, the following are obtained.

$$\mathrm{Var}\left[\hat{P}_c(i)\right] = \frac{P_c(i)[1 - P_c(i)]}{N} \tag{2-15}$$

$$\mathrm{Cov}\left[\hat{P}_c(i)\hat{P}_c(j)\right] = -\frac{P_c(i)P_c(j)}{N} \tag{2-16}$$

$$\mathrm{Var}\left(\hat{\lambda}_{ii}\right) = \frac{\lambda_{ii}(1 - \lambda_{ii})}{mP_c(i)} \tag{2-17}$$

$$\mathrm{Cov}\left[\hat{\lambda}_{ii}\hat{P}_c(j)\right] = \mathrm{Cov}\left[\hat{P}_c(i)\hat{\lambda}_{jj}\right] = \mathrm{Cov}\left(\hat{\lambda}_{ii}\hat{\lambda}_{kk}\right) = 0 \tag{2-18}$$

for all i and j, $i \neq k$, where

$$N = \sum_{\ell=1}^{M} X_\ell \tag{2-19}$$

Substituting equations (2-5) through (2-19) into equation (2-14) yields an expression for the $\mathrm{Var}\left(\hat{P}_{cc}\right)$ as follows.

$$\mathrm{Var}\left(\hat{P}_{cc}\right) = \sum_{i=1}^{M} \frac{\lambda_{ii}(1 - \lambda_{ii})}{mP_c(i)} P_c^2(i) + \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} \frac{[-P_c(i)P_c(j)]}{N} \lambda_{ii}\lambda_{jj}$$

$$+ \sum_{i=1}^{M} \frac{P_c(i)[1 - P_c(i)]}{N} \lambda_{ii}^2$$

$$= \sum_{i=1}^{M} \frac{\lambda_{ii}(1 - \lambda_{ii})}{m} P_c(i) + \sum_{i=1}^{M} \frac{P_c(i)\lambda_{ii}^2}{N} - \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{P_c(i)P_c(j)\lambda_{ii}\lambda_{jj}}{N}$$

$$= \sum_{i=1}^{M} \frac{\lambda_{ii}(1 - \lambda_{ii})P_c(i)}{m} + \frac{\left[\sum_{i=1}^{M} P_c(i)\lambda_{ii}^2 - P_{cc}^2\right]}{N} \tag{2-20}$$

Following a similar analysis, an expression may be obtained for the asymptotic variance of the a priori probability estimator (ref. 15).

$$\mathrm{Var}\left(\hat{P}_i\right) = \sum_{j=1}^{M} \frac{\lambda_{ij}(1 - \lambda_{ij})P_c(j)}{m} + \frac{\left[\sum_{j=1}^{M} P_c(j)\lambda_{ij}^2 - P_c^2(i)\right]}{N} \qquad (2\text{-}21)$$

In general, one can obtain expressions for sample sizes m and N, either by minimizing the $\mathrm{Var}\left(\hat{P}_{cc}\right)$ or by minimizing the $\mathrm{Var}\left(\hat{P}_i\right)$, subject to some cost constraints.

3. MAXIMUM LIKELIHOOD ESTIMATION WITH LABEL IMPERFECTIONS

In practical situations, obtaining labels is expensive, and very often these labels are imperfect. In this section, we formulate the problem of estimating, with imperfections in the labels, the various quantities considered in section 2.

It is assumed that the classifier is trained on representative data, and a set of labeled patterns (possibly with imperfect labels) and a set of unlabeled patterns are presented to the classifier. The classifier classifies these patterns, and the results are matrices similar to table 2-1. Now the various quantities are defined as follows.

Let ω' be the imperfect label, $P'_i = P(\omega' = i)$ be the a priori probability that the imperfect label is i, $p'_{ij} = P(\omega' = i, \omega_c = j)$ be the probability that the imperfect label is i, and j be the classifier label. Consider

$$
\begin{aligned}
p'_{ij} &= P(\omega' = i, \omega_c = j) \\
&= \sum_{\ell=1}^{M} P(\omega' = i, \omega = \ell, \omega_c = j) \\
&= \sum_{\ell=1}^{M} P(\omega' = i \,|\, \omega = \ell, \omega_c = j) P(\omega = \ell, \omega_c = j) \\
&= \sum_{\ell=1}^{M} P(\omega' = i \,|\, \omega = \ell) P(\omega_c = j \,|\, \omega = \ell) P(\omega = \ell) \qquad (3\text{-}1)
\end{aligned}
$$

where it is assumed that

$$
P(\omega' = i \,|\, \omega = \ell) = P(\omega' = i \,|\, \omega = \ell, \omega_c = j) \qquad (3\text{-}2)
$$

This assumption states that, given the true label and the classifier label, the imperfect label depends only on the true label. This is a reasonable assumption. In acquiring the label for a pattern, the labeler depends heavily on the true label of the pattern and virtually does not know the

classifier label. (In labeling a pixel in imagery data, the assigned label depends on the true label of the pixel and its neighbors and on some other data such as ancillary information.) Now consider

$$P_c(j) = P(\omega_c = j)$$

$$= \sum_{\ell=1}^{M} P(\omega_c = j, \omega = \ell)$$

$$= \sum_{\ell=1}^{M} P(\omega_c = j | \omega = \ell) P(\omega = \ell) \qquad (3\text{-}3)$$

Substituting equations (3-1) and (3-3) into the likelihood function and taking the logarithm results in

$$L = \log C + \sum_{i=1}^{M} \sum_{j=1}^{M} m_{ij} \log \left[\sum_{\ell=1}^{M} P(\omega' = i | \omega = \ell) P(\omega_c = j | \omega = \ell) P(\omega = \ell) \right]$$

$$+ \sum_{j=1}^{M} X_j \log \left[\sum_{\ell=1}^{M} P(\omega_c = j | \omega = \ell) P(\omega = \ell) \right] \qquad (3\text{-}4)$$

Finding closed-form solutions for the parameters by maximizing L seems to be difficult, since the resulting equations become coupled in terms of parameters. However, optimization techniques, such as the Davidon-Fletcher-Powell procedure, can be used to maximize L (refs. 16-18). Now, the problem can be formulated as

Find: $P(\omega' = i | \omega = \ell), P(\omega_c = j | \omega = \ell), P(\omega = \ell)$; $i,j,\ell = 1,2,\cdots,M$

such that L is maximized subject to the following constraints.

$$
\begin{aligned}
&\sum_{i=1}^{M} P(\omega' = i \mid \omega = \ell) = 1 \quad ; \quad \ell = 1,2,\cdots,M \\
&\sum_{j=1}^{M} P(\omega_c = j \mid \omega = \ell) = 1 \quad ; \quad \ell = 1,2,\cdots,M \\
&\sum_{\ell=1}^{M} P(\omega = \ell) = 1 \\
&P(\omega' = i \mid \omega = \ell) \geq 0 \quad ; \quad i,\ell = 1,2,\cdots,M \\
&P(\omega_c = j \mid \omega = \ell) \geq 0 \quad ; \quad j,\ell = 1,2,\cdots,M \\
&P(\omega = \ell) \geq 0 \quad ; \quad \ell = 1,2,\cdots,M
\end{aligned}
\tag{3-5}
$$

The numbers of parameters and constraints for different values of M are listed in table 3-1.

TABLE 3-1.— PARAMETERS AND CONSTRAINTS FOR A GENERAL CASE

Number of classes, M	Number of parameters, $2M^2+M$	Number of constraints	
		Equality, $2M+1$	Inequality, $2M^2+M$
2	10	5	10
3	21	7	21
4	36	9	36
5	55	11	55

As indicated in table 3-1, the numbers of parameters and constraints increase with the square of the number of classes, resulting in a large number of degrees of freedom for the optimization problem. However, the numbers of constraints and parameters can be reduced by modeling the label imperfections and the probabilities of misclassification.

3.1 MAXIMUM LIKELIHOOD ESTIMATION WITH SIMPLIFIED MODELS

This section provides (1) models for label imperfections and probabilities of misclassification and (2) a formulation of the problem of maximum likelihood estimation. To develop a model for describing the probabilities of imperfections in the labels, consider the following.

a. If there are no imperfections in the labels, for different i and j,

$$\left.\begin{aligned} P(\omega' = i|\omega = i) &= 1 \\ \text{and} \quad P(\omega' = j|\omega = i) &= 0 \end{aligned}\right\} \tag{3-6}$$

b. If the imperfect label for a pattern is assigned purely at random, irrespective of its true label, for different i and j,

$$\left.\begin{aligned} P(\omega' = i|\omega = i) &= \frac{1}{M} \\ \text{and} \quad P(\omega' = j|\omega = i) &= \frac{1}{M} \end{aligned}\right\} \tag{3-7}$$

Since, in a practical situation, the assignment of a label lies somewhere between the above two extremes, the imperfections in the labels can be modeled through a parameter θ_1, which lies between 0 and 1 as

$$\left.\begin{aligned} P(\omega' = i|\omega = i) &= \frac{(1 - \theta_1)}{M} + \theta_1 \\ P(\omega' = j|\omega = i) &= \frac{(1 - \theta_1)}{M} \end{aligned}\right\} \tag{3-8}$$

where $0 \leq \theta_1 \leq 1$.

From equations (3-6) through (3-8), it is easily seen that $\theta_1 = 1$ denotes no imperfections in the labels and $\theta_1 = 0$ denotes random labeling. The following shows that this definition satisfies the postulates of probability. Consider the following.

$$\sum_{j=1}^{M} P(\omega' = j|\omega = i) = P(\omega' = i|\omega = i) + \sum_{\substack{j=1 \\ j\neq i}}^{M} P(\omega' = j|\omega = i)$$

$$= \frac{(1 - \theta_1)}{M} + \theta_1 + \sum_{\substack{j=1 \\ j\neq i}}^{M} \frac{(1 - \theta_1)}{M} = \theta_1 + \overline{1 - \theta_1} = 1 \qquad (3\text{-}9)$$

thus satisfying the probability rule. However, it is noted that the imperfections in the labels can be modeled through some other parameter; for example, making $\theta = \frac{\alpha}{1 + \alpha}$ causes the imperfections to be dependent on α, $0 \leq \alpha \leq \infty$; or, making $\theta = \frac{e^{-\beta}}{1 + e^{-\beta}}$ causes the imperfections to be dependent on β, $-\infty \leq \beta \leq \infty$. In this section, it is assumed that the imperfections are modeled through equation (3-8).

Similarly, classification errors can be modeled as follows

a. If there are no classification errors, for different i and j,

$$\left.\begin{aligned} P(\omega_c = i|\omega = i) &= 1 \\ \text{and} \quad P(\omega_c = j|\omega = i) &= 0 \end{aligned}\right\} \qquad (3\text{-}10)$$

b. If the classifier is making random decisions, for different i and j,

$$\left.\begin{aligned} P(\omega_c = i|\omega = i) &= \frac{1}{M} \\ \text{and} \quad P(\omega_c = j|\omega = i) &= \frac{1}{M} \end{aligned}\right\} \qquad (3\text{-}11)$$

Since, in general, the truth lies somewhere between the above two extremes, the classification errors can be modeled through a parameter θ_2, which lies between 0 and 1 as

$$\left.\begin{aligned} P(\omega_c = i|\omega = i) &= \frac{(1 - \theta_2)}{M} + \theta_2 \\ \text{and} \quad P(\omega_c = j|\omega = i) &= \frac{(1 - \theta_2)}{M} \end{aligned}\right\} \qquad (3\text{-}12)$$

where $0 \leq \theta_2 \leq 1$. As before, it can be seen that this model satisfies the postulates of probability.

Let $\lambda_1 = (1 - \theta_1)$ and $\lambda_2 = \theta_1$; then $\lambda_1 + \lambda_2 = 1$. Similarly, let $\lambda_3 = (1 - \theta_2)$ and $\lambda_4 = \theta_2$; then $\lambda_3 + \lambda_4 = 1$. The following expresses the likelihood function in terms of the above models. Consider

$$P_c(j) = P(\omega_c = j)$$

$$= \sum_{\ell=1}^{M} P(\omega = \ell)P(\omega_c = j|\omega = \ell)$$

$$= P(\omega_c = j|\omega = j)P(\omega = j) + \sum_{\substack{\ell=1 \\ \ell \neq j}}^{M} P(\omega = \ell)P(\omega_c = j|\omega = \ell)$$

$$= \left(\frac{\lambda_3}{M} + \lambda_4\right)P_j + \sum_{\substack{\ell=1 \\ \ell \neq j}}^{M} \frac{\lambda_3}{M} P_\ell = \frac{\lambda_3}{M} + \lambda_4 P_j \qquad (3\text{-}13)$$

$$P'_{ii} = P(\omega' = i, \omega_c = i)$$

$$= \sum_{\ell=1}^{M} P(\omega' = i|\omega = \ell)P(\omega_c = i|\omega = \ell)P(\omega = \ell)$$

$$= \sum_{\substack{\ell=1 \\ \ell \neq i}}^{M} P(\omega' = i|\omega = \ell)P(\omega_c = i|\omega = \ell)P(\omega = \ell)$$

$$+ P(\omega' = i|\omega = i)P(\omega_c = i|\omega = i)P(\omega = i)$$

$$= \sum_{\substack{\ell=1 \\ \ell \neq i}}^{M} \frac{\lambda_1}{M}\frac{\lambda_3}{M} P_\ell + \left(\frac{\lambda_1}{M} + \lambda_2\right)\left(\frac{\lambda_3}{M} + \lambda_4\right)P_i$$

$$= \frac{\lambda_1}{M}\frac{\lambda_3}{M} + (\lambda_1\lambda_4 + \lambda_2\lambda_3)\frac{P_i}{M} + \lambda_2\lambda_4 P_i \qquad (3\text{-}14)$$

Similarly, for $i \neq j$,

$$p'_{ij} = P(\omega' = i, \omega_c = j)$$

$$= \sum_{\ell=1}^{M} P(\omega' = i \mid \omega = \ell) P(\omega_c = j \mid \omega = \ell) P(\omega = \ell)$$

$$= \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^{M} P(\omega' = i \mid \omega = \ell) P(\omega_c = j \mid \omega = \ell) P(\omega = \ell)$$

$$+ P(\omega' = i \mid \omega = i) P(\omega_c = j \mid \omega = i) P(\omega = i)$$

$$+ P(\omega' = i \mid \omega = j) P(\omega_c = j \mid \omega = j) P(\omega = j)$$

$$= \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^{M} \frac{\lambda_1}{M} \frac{\lambda_3}{M} P_\ell + \left(\frac{\lambda_1}{M} + \lambda_2\right) \frac{\lambda_3}{M} P_i + \frac{\lambda_1}{M} \left(\frac{\lambda_3}{M} + \lambda_4\right) P_j$$

$$= \frac{\lambda_1}{M} \frac{\lambda_3}{M} + \frac{\lambda_2 \lambda_3}{M} P_i + \frac{\lambda_1 \lambda_4}{M} P_j \qquad (3\text{-}15)$$

Substituting equations (3-13) through (3-15) into the likelihood function results in

$$L = \log C + \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} m_{ij} \log\left(\frac{\lambda_1 \lambda_3}{M^2} + \frac{\lambda_2 \lambda_3}{M} P_i + \frac{\lambda_1 \lambda_4}{M} P_j\right)$$

$$+ \sum_{i=1}^{M} m_{ii} \log\left[\frac{\lambda_1 \lambda_3}{M^2} + \left(\frac{\lambda_1 \lambda_4}{M} + \frac{\lambda_2 \lambda_3}{M} + \lambda_2 \lambda_4\right) P_i\right]$$

$$+ \sum_{i=1}^{M} \chi_i \log\left(\frac{\lambda_3}{M} + \lambda_4 P_i\right) \qquad (3\text{-}16)$$

Now, the problem can be stated as follows.

Find: λ_i (i = 1,2,3,4) and P_j (j = 1,2,•••,M)

so that L is maximized subject to the following constraints.

$$\left.\begin{aligned} \sum_{i=1}^{M} P_i &= 1 \\ \lambda_1 + \lambda_2 &= 1 \\ \lambda_3 + \lambda_4 &= 1 \\ \lambda_i \geq 0 \quad ; \quad i &= 1,\cdots,4 \\ P_i \geq 0 \quad ; \quad i &= 1,2,\cdots,M \end{aligned}\right\} \tag{3-17}$$

Optimization techniques, such as the Davidon-Fletcher-Powell procedure, can be used to maximize L (refs. 16-18). The numbers of parameters and constraints for different values of M are listed in table 3-2.

TABLE 3-2.— PARAMETERS AND CONSTRAINTS FOR A SIMPLIFIED PROBLEM

Number of classes, M	Number of parameters, 4+M	Number of constraints	
		Equality, 3	Inequality, 4+M
2	6	3	6
3	7	3	7
4	8	3	8
5	9	3	9

Table 3-2 indicates that the optimization problem is considerably simplified.

3.2 A PRACTICAL APPLICATION

The maximum likelihood estimation with the simplified models presented in section 3.1 is applied to processing remotely sensed Landsat multispectral scanner (MSS) data. Several segments[1] are processed in the following manner. A linear classifier is trained for two classes. Class 1 is wheat (W) and class 2 is other (N). This classifier is used to classify a test set of data (104 patterns) for which labels are available and a set of data (209 patterns) for which labels are not available. Thus, the classifications corresponding to table 2-1 are computed. The labels for the test data are assumed to be imperfect. The maximum likelihood estimates of λ_i ($i = 1,2,3,4$) and P_j ($j = 1,2$), subject to the constraints of equation (3-17), are obtained using the Davidon-Fletcher-Powell optimization procedure (refs. 16,17).

The Davidon-Fletcher-Powell procedure, in conjunction with an exterior penalty function, very efficiently carries out the optimization of the performance function, subject to various constraints. In general, these constraints must be continuous differentiable functions of the parameters. The original likelihood function is augmented with the functions of the constraints. The augmented likelihood function is penalized whenever the constraints are violated. For sufficiently large penalties, the unconstrained optimization of the augmented likelihood function can be shown to be equivalent to the original constrained optimization.

The results obtained from the optimization of the likelihood function are shown in table 3-3. The last column in table 3-3 lists the $P(\omega = 1)$ values computed from the ground-truth information over the entire segment for each segment. The following conclusions can be made from table 3-3. The mean and variance of errors of estimated P_1 with respect to the ground-truth P_1 are smaller with the modeling of imperfections in the labels than with the

[1] A segment is a 9- by 11-kilometer (5- by 6-nautical mile) area for which the MSS image is divided into a rectangular array of pixels, 117 rows by 196 columns.

TABLE 3-3.— ESTIMATES OF A PRIORI PROBABILITY AND P_{cc} WITH AND WITHOUT MODELING OF IMPERFECTIONS IN THE LABELS

Segment	Site description		Without modeling imperfections in the labels		With modeling imperfections in the labels			Ground-truth proportion, $P(\omega=1)$
	County	State	P_1'	P_{cc}'	$P(\omega'=1 \mid \omega=1)$ (a)	P_{cc}	$P_1=P(\omega=1)$ (b)	
1060	Sherman	Tex.	0.3421	0.8284	0.8377	0.9905	0.2492	0.229
1512	Clay	Minn.	.4295	.7653	.7678	1.0000	.3594	.337
1520	Big Stone	Minn.	.2647	.7763	1.0000	.7790	.2759	.299
1604	Renville	N. Dak.	.5506	.6378	.7100	.8363	.6030	.526
1648	Spink	S. Dak.	.2868	.8160	1.0000	.8182	.2894	.379
1677	Spink	S. Dak.	.3838	.7501	.7847	.9445	.3034	.341
1734	Hill	Mont.	.4663	.8857	.8865	1.0000	.4486	.440
1929	Blaine	Mont.	.4445	.9422	1.0000	.9472	.4672	.426
Mean of errors			0.02391				0.002388	
Variance of errors			0.00374				0.002318	

[a]Probability of label imperfections.

[b]Estimated proportion of class 1.

estimates obtained assuming the labels are perfect. When there are no imperfections in the labels (i.e., for segments 1520, 1648, and 1929), the estimates of P_{cc}'s obtained with and without modeling of imperfections in the labels are identical. Furthermore, when the estimated P_{cc} is 1 (with modeling of label imperfections), the estimated P_{cc} (assuming labels are perfect) is identical with the probability of label imperfections. The P_1 and P_1' are related as follows

$$P_1' = P(\omega' = 1) = \sum_{\ell=1}^{M} P(\omega' = 1 | \omega = \ell) P(\omega = \ell) \qquad (3\text{-}18)$$

If it is assumed that the labels are perfect, the estimate of P_1 is an estimate of P_1'. Table 3-4 lists the estimate of P_1' obtained from equation (3-18) and that obtained as a maximum likelihood estimate from equation (2-10), assuming the labels are perfect.

TABLE 3-4.— COMPARISON OF ESTIMATES OF P_1 WITH AND WITHOUT MODELING OF LABEL IMPERFECTIONS

Segment	Estimate of P_1', $P_1' = \sum_{j=1}^{M} P(\omega'=1 \mid \omega=j) P(\omega=j)$	Maximum likelihood estimate of P_1' obtained from equation (2-10)
1060	0.3322	0.3421
1512	.4246	.4295
1520	.2759	.2647
1604	.5432	.5506
1648	.2894	.2868
1677	.3880	.3838
1734	.4602	.4663
1929	.4672	.4445

Columns 2 and 3 of table 3-4 are almost identical, thus verifying the validity of the models used in defining the label imperfections.

3.3 MAXIMUM LIKELIHOOD ESTIMATION WITH CLASS-DEPENDENT MODELING OF LABEL IMPERFECTIONS AND ERROR PROBABILITIES

When modeling label imperfections and error probabilities, the θ's and hence λ's can be made class dependent, which increases the complexity of the problem. For different i and j, the imperfections in the labels can be modeled as

$$\left.\begin{aligned} P(\omega' = i|\omega = i) &= \frac{[1 - \theta_1(i)]}{M} + \theta_1(i) \\ P(\omega' = j|\omega = i) &= \frac{[1 - \theta_1(i)]}{M} \\ 0 \le \theta_1(i) &\le 1 \end{aligned}\right\} \qquad (3\text{-}19)$$

Similarly, for different i and j, the error probabilities can be modeled as

$$\left.\begin{aligned} P(\omega_c = i|\omega = i) &= \frac{[1 - \theta_2(i)]}{M} + \theta_2(i) \\ P(\omega_c = j|\omega = i) &= \frac{[1 - \theta_2(i)]}{M} \\ 0 \le \theta_2(i) &\le 1 \end{aligned}\right\} \qquad (3\text{-}20)$$

It can be shown that these models satisfy the postulates of probability. Let $\lambda_1(i) = [1 - \theta_1(i)]$, $\lambda_2(i) = \theta_1(i)$, $\lambda_3(i) = [1 - \theta_2(i)]$, and $\lambda_4(i) = \theta_2(i)$. Then,

$$\left.\begin{aligned} \lambda_1(i) + \lambda_2(i) &= 1 \quad ; \quad i = 1,2,\cdots,M \\ \lambda_3(i) + \lambda_4(i) &= 1 \end{aligned}\right\} \qquad (3\text{-}21)$$

An analysis similar to equations (3-13) through (3-15) yields the following equations.

$$\begin{aligned} P_c(j) &= P(\omega_c = j) \\ &= \sum_{\ell=1}^{M} \frac{\lambda_3(\ell)}{M} P_\ell + \lambda_4(j) P_j \end{aligned} \qquad (3\text{-}22)$$

$$p'_{ii} = P(\omega' = i, \omega_c = i)$$

$$= \sum_{\ell=1}^{M} \frac{\lambda_1(\ell)}{M} \frac{\lambda_3(\ell)}{M} P_\ell + \left[\frac{\lambda_1(i)}{M} \lambda_4(i) + \lambda_2(i) \frac{\lambda_3(i)}{M} + \lambda_2(i)\lambda_4(i) \right] P_i \tag{3-23}$$

$$p'_{ij} = P(\omega' = i, \omega_c = j)$$

$$= \sum_{\ell=1}^{M} \frac{\lambda_1(\ell)}{M} \frac{\lambda_3(\ell)}{M} P_\ell + \left[\lambda_2(i) \frac{\lambda_3(i)}{M} P_i + \frac{\lambda_1(j)}{M} \lambda_4(j) P_j \right] \tag{3-24}$$

Equations (3-22) through (3-24) can be used to express the likelihood function as follows.

$$L = \log C + \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} m_{ij} \log \left\{ \sum_{\ell=1}^{M} \frac{\lambda_1(\ell)\lambda_3(\ell)}{M} P_\ell + \left[\frac{\lambda_2(i)\lambda_3(i)}{M} P_i + \frac{\lambda_1(j)}{M} \lambda_4(j) P_j \right] \right\}$$

$$+ \sum_{i=1}^{M} m_{ii} \log \left\{ \sum_{\ell=1}^{M} \frac{\lambda_1(\ell)}{M} \frac{\lambda_3(\ell)}{M} P_\ell + \left[\frac{\lambda_1(i)}{M} \lambda_4(i) + \frac{\lambda_2(i)\lambda_3(i)}{M} + \lambda_2(i)\lambda_4(i) \right] P_i \right\}$$

$$+ \sum_{i=1}^{M} X_i \log \left[\sum_{\ell=1}^{M} \frac{\lambda_3(\ell)}{M} P_\ell + \lambda_4(i) P_i \right] \tag{3-25}$$

The problem of maximizing L may be stated as follows:

Find: $\lambda_i(j)$ $(j = 1,2,\cdots,M;\ i = 1,2,3,4)$ and P_j $(j = 1,2,\cdots,M)$

so that L is maximized subject to the following constraints.

$$\left.\begin{aligned} \sum_{i=1}^{M} P_i &= 1 \\ \lambda_1(i) + \lambda_2(i) &= 1 \quad ; \quad i = 1,2,\cdots,M \\ \lambda_3(i) + \lambda_4(i) &= 1 \quad ; \quad i = 1,2,\cdots,M \\ \lambda_i(j) &\geq 0 \quad ; \quad i = 1,2,3,4 \text{ and } j = 1,2,\cdots,M \\ P_i &\geq 0 \quad ; \quad i = 1,2,\cdots,M \end{aligned}\right\} \tag{3-26}$$

The optimization technique of Davidon, Fletcher, and Powell (refs. 16,17) can be used to maximize L in equation (3-25), subject to the constraints of equation (3-26). The numbers of parameters and constraints for different values of M are listed in table 3-5.

TABLE 3-5.— PARAMETERS AND CONSTRAINTS FOR CLASS-DEPENDENT MODELS

Number of classes, M	Number of parameters, 4M+M	Number of constraints	
		Equality, 2M+1	Inequality, 4M+M
2	10	5	10
3	15	7	15
4	20	9	20
5	25	11	25

Table 3-5 shows that the numbers of parameters and constraints grow linearly with M.

4. IDENTIFICATION OF MISLABELED PATTERNS

This section considers the problem of identifying mislabeled patterns, if the probability of label imperfections is either known or estimated using the methods developed in section 3. Some relationships are developed between the a priori probabilities and the probability densities with and without imperfections in the labels. The imperfections in the labels are described by the probabilities

$$\beta_{ji} = P(\omega' = i|\omega = j) \quad ; \quad i,j = 1,2,\cdots,M \tag{4-1}$$

where i and j indicate class. We have the constraint,

$$\sum_{i=1}^{M} \beta_{ji} = 1 \tag{4-2}$$

It is assumed that

$$p(X|\omega = j) = p(X|\omega' = i,\omega = j) \tag{4-3}$$

That is, given the true label of a pattern, the density of the pattern does not depend on its imperfect label. To obtain the relationship between $p(X|\omega = i)$ and $p(X|\omega' = i)$, consider

$$\begin{aligned} p(X|\omega' = i) &= \frac{1}{P(\omega' = i)} \sum_{j=1}^{M} p(X,\omega' = i,\omega = j) \\ &= \frac{1}{P(\omega' = i)} \sum_{j=1}^{M} p(X|\omega' = i,\omega = j)P(\omega' = i|\omega = j)P(\omega = j) \\ &= \frac{1}{P(\omega' = i)} \sum_{j=1}^{M} \beta_{ji}P(\omega = j)p(X|\omega = j) \end{aligned} \tag{4-4}$$

Similarly, the a priori probabilities are related as

$$P(\omega' = i) = \sum_{j=1}^{M} \beta_{ji}P(\omega = j) \tag{4-5}$$

Inverting equation (4-4) yields the following result for the two-class case.

$$\left.\begin{aligned} P(\omega = 1)p(X|\omega = 1) &= \frac{1}{(\beta_{11}\beta_{22} - \beta_{12}\beta_{21})}\,[\beta_{22}P(\omega' = 1)p(X|\omega' = 1) \\ &\quad - \beta_{21}P(\omega' = 2)p(X|\omega' = 2)] \\ P(\omega = 2)p(X|\omega = 2) &= \frac{1}{(\beta_{11}\beta_{22} - \beta_{12}\beta_{21})}\,[\beta_{11}P(\omega' = 2)p(X|\omega' = 2) \\ &\quad - \beta_{12}P(\omega' = 1)p(X|\omega' = 1)] \end{aligned}\right\} \qquad (4\text{-}6)$$

Let

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2M} \\ \vdots & \vdots & & \vdots \\ \beta_{M1} & \beta_{M2} & \cdots & \beta_{MM} \end{bmatrix} \qquad (4\text{-}7)$$

Assuming β^{-1} exists, the following can be obtained from equation (4-4) in the multiclass case.

$$P(\omega = i)p(X|\omega = i) = \sum_{s=1}^{M} \delta_{is}P(\omega' = s)p(X|\omega' = s) \quad ; \quad i = 1,2,\cdots,M \qquad (4\text{-}8)$$

4.1 IDENTIFICATION OF MISLABELED PATTERNS IN THE TWO-CLASS CASE

The following expressions are developed for the identification of mislabeled patterns using a linear classifier. The linear classifier implements a decision criterion

$$\left.\begin{aligned} &\text{Decide } X \in \omega' = 1 \text{ if } g(X) = W^T X + w_0 > 0 \\ &\text{Decide } X \in \omega' = 2 \text{ otherwise} \end{aligned}\right\} \qquad (4\text{-}9)$$

It is assumed that $p(X|\omega' = i)$ is multivariate normal; i.e., $p(X|\omega' = i) \sim N(M_i', \Sigma_i')$, $i = 1,2$. Since $g(X)$ is a linear combination of the components of pattern vector X, if X is normally distributed, $g(X)$ is also normally distributed. That is,

$$p[g(X)|X \in \omega' = i] \sim N\left[m_i', (\sigma_i')^2\right] \quad ; \quad i = 1,2 . \tag{4-10}$$

where

$$\left.\begin{aligned} m_i' &= W^T M_i' + w_0 \\ (\sigma_i')^2 &= W^T \Sigma_i' W \end{aligned}\right\} \tag{4-11}$$

To identify and change the labels of mislabeled patterns, the following scheme is proposed.

$$\left.\begin{aligned} &\text{Change the label of X to } \omega = 1 \text{ if } g(X) > t_1 \\ &\text{Change the label of X to } \omega = 2 \text{ if } g(X) < -t_2 \\ &\text{Do not change the label of X if } -t_2 \le g(X) \le t_1 \end{aligned}\right\} \tag{4-12}$$

The thresholds t_1 and $-t_2$ are used to identify the incorrect labels and are determined by specifying the probability α, that mislabeling will occur in the label correction process. An expression for the probability that the label correction scheme will give an incorrect label is derived in the following equation.

$$\begin{aligned} P_{BL} &= P(\text{bad label}) \\ &= P(\omega = 1)P(\text{bad label}|X \in \omega = 1) + P(\omega = 2)P(\text{bad label}|X \in \omega = 2) \\ &= P(\omega = 1)P[g(X) < -t_2|X \in \omega = 1] + P(\omega = 2)P[g(X) > t_1|X \in \omega = 2] \end{aligned} \tag{4-13}$$

Using equations (4-6) and (4-13) obtains the following result.

$$P(\omega = 1)P[g(X) < -t_2 | X \in \omega = 1] = P(\omega = 1)\int_{-\infty}^{-t_2} p[g(X)|X \in \omega = 1]d[g(X)]$$

$$= \frac{1}{(\beta_{11}\beta_{22} - \beta_{12}\beta_{21})}\left\{\beta_{22}P(\omega' = 1)\int_{-\infty}^{-t_2} p[g(X)|\omega' = 1]d[g(X)]\right.$$

$$\left. - \beta_{21}P(\omega' = 2)\int_{-\infty}^{-t_2} p[g(X)|\omega' = 2]d[g(X)]\right\}$$

$$= \frac{1}{(\beta_{11}\beta_{22} - \beta_{12}\beta_{21})}\left[\beta_{22}P(\omega' = 1)\int_{-\infty}^{\frac{-t_2-m_1'}{\sigma_1'}} \psi(y)dy\right.$$

$$\left. - \beta_{21}P(\omega' = 2)\int_{-\infty}^{\frac{-t_2-m_2'}{\sigma_2'}} \psi(y)dy\right] \qquad (4\text{-}14)$$

where

$$\psi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right) \qquad (4\text{-}15)$$

Similarly,

$$P(\omega = 2)P[g(X) > t_1 | X \in \omega = 2] = \frac{1}{(\beta_{11}\beta_{22} - \beta_{12}\beta_{21})}\left[\beta_{11}P(\omega' = 2)\right.$$

$$\left.\int_{-\infty}^{\frac{-t_1+m_2'}{\sigma_2'}} \psi(y)dy - \beta_{12}P(\omega' = 1)\int_{-\infty}^{\frac{-t_1+m_1'}{\sigma_1'}} \psi(y)dy\right] \qquad (4\text{-}16)$$

From equations (4-13) through (4-16), the probability of a bad label P_{BL} can be obtained as

$$P_{BL} = \frac{1}{(\beta_{11}\beta_{22} - \beta_{12}\beta_{21})} \left\{ \left[\beta_{22} P(\omega' = 1) \int_{-\infty}^{\frac{-t_2 - m_1'}{\sigma_1'}} \psi(y)dy \right. \right.$$

$$\left. - \beta_{21} P(\omega' = 2) \int_{-\infty}^{\frac{-t_2 - m_2'}{\sigma_2'}} \psi(y)dy \right] + \left[\beta_{11} P(\omega' = 2) \int_{-\infty}^{\frac{-t_1 + m_2'}{\sigma_2'}} \psi(y)dy \right.$$

$$\left. \left. - \beta_{12} P(\omega' = 1) \int_{-\infty}^{\frac{-t_1 + m_1'}{\sigma_1'}} \psi(y)dy \right] \right\} \qquad (4\text{-}17)$$

For a given α, t_1 and $-t_2$ can be computed using an optimization technique such as the Davidon-Fletcher-Powell procedure, so that the square of the error between α and P_{BL} is minimized and can be used in the incorrect label identification scheme.

4.2 AN EXAMPLE OF APPLICATION OF THE INCORRECT LABEL IDENTIFICATION SCHEME

The two-class imperfect label correction scheme presented in section 4.1 is applied to a practical problem in remote sensing. In particular, it is applied to Landsat imagery of segment 1060. Data from two acquisitions are processed, and each acquisition has four spectral bands. The image is overlaid with a rectangular grid of 209 grid intersections, and the labels of pixels corresponding to each grid intersection are acquired. A linear classifier is trained on one-half of the data. The remaining one-half of the data is used as a test data set. Test data set and total data set classifications are obtained using the linear classifier. This results in matrices corresponding to table 2-1(a) and (b). The maximum likelihood estimates of

label inperfections are obtained using the simplified models presented in section 3.1. The β-matrix and the a priori probabilities obtained are

$$\left.\begin{aligned} \beta &= \begin{bmatrix} 0.8378 & 0.1622 \\ 0.1622 & 0.8378 \end{bmatrix} \\ P(\omega = 1) &= 0.24921 \\ P(\omega = 2) &= 0.75079 \end{aligned}\right\} \qquad (4\text{-}18)$$

If $\alpha = 0.001$ is chosen, upper and lower thresholds t_1 and $-t_2$ that minimize the square of the difference between α and P_{BL} are computed using the Davidon-Fletcher-Powell procedure. The patterns of class $\omega' = 2$, the discriminant function values of which exceeded t_1, and the patterns of class $\omega' = 1$, the discriminant function values of which are less than $-t_2$, are identified and marked with circles in figures 4-1 and 4-2. These figures list the labels of the pixels of 209 grid intersections and their relative positions.

Films of the two acquisitions of segment 1060 used in the processing were examined by an analyst-interpreter (AI), and the results are given in figures 4-3 and 4-4.

From an analysis of figures 4-3 and 4-4, it can be concluded that the decisions of the label correction scheme are in close agreement with the AI interpretations of the imagery films.

4.3 IDENTIFICATION OF MISLABELED PATTERNS IN THE MULTICLASS CASE

Let $g_i(X)$ be the discriminant function of the ith class $\omega' = i$, where

$$g_i(X) = W_i^T X + w_{i0} \quad ; \quad i = 1,2,\cdots,M \qquad (4\text{-}19)$$

The usual decision criterion in a multiclass case is to decide $X \in \omega' = \ell$, if

$$g_\ell(X) = \max_{\substack{j \\ j=1,2,\cdots,M \\ j \neq \ell}} g_j(X) \qquad (4\text{-}20)$$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	N	(N)*	N	N	N	N	N	N	N	N	N		N			N	N	N	N
2	N	N	N	N			N	N	N	N	N				N	N	N	N	N
3	N	N	N	N	N	N	N	N	N	N	N	N		N	(N)B	N			
4	N	N	N	N	N		N	N		N	N	N	N	N		(N)			
5	N	N	N	N	N						N	N	N	N	N	N	N	(N)	
6	N	N	N	N	N	N						N	N	N			N	N	N
7	N	N		N	N	N	N			N	N	N	N	N	N	N		N	N
8	N		N	N	N				N			N	N	N	N		N	N	N
9	N			N		N	N	N	N	N	N							N	
10		N	N	N			N	N			(N)				N	N		N	N
11					(N)		N	N	N	N	N		N	N	N	N	N	N	N

Computed upper threshold $t_1 = 0.1507$

Legend	
Blank	Wheat pixels
N	Other pixels
(N)	Pixels identified by label correction scheme as wheat
B	AI decision as wheat but bordering class other
*	AI decision as other

Figure 4-1.— Diagram of 209 grid intersections showing pixels labeled other and other pixels reidentified as wheat using imperfect label identification scheme.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1												W		(W)	W				
2					(W)	(W)						(W)	(W)$_B$	(W)					
3													W				W	(W)	(W)
4						W			W						(W)		W	W	W
5						W	(W)$_B$	W	W	W									W
6							(W)	W	W	W	W				W	(W)			
7			(W)					W	(W)*								(W)		
8		(W)				W	W	W		W	W					(W)			
9		(W)	W		W							W	W	W	(W)$_B$	W	W		(W)
10	(W)				W	W			W	W		W	W	W			(W)		
11	W	W	(W)*	W		W						W							

Computed lower threshold $-t_2 = -0.01628$

Legend	
Blank	Other pixels
W	Wheat pixels
(W)	Pixels identified by label correction scheme as other
B	AI decision as other but bordering wheat
*	AI decision as wheat

Figure 4-2.— Diagram of 209 grid intersections showing pixels labeled wheat and wheat pixels identified as other using imperfect label identification scheme.

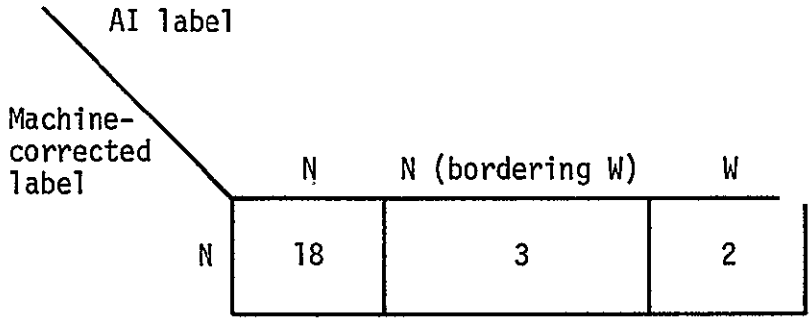

Figure 4-3.— AI labels for patterns where labels were changed from wheat to other.

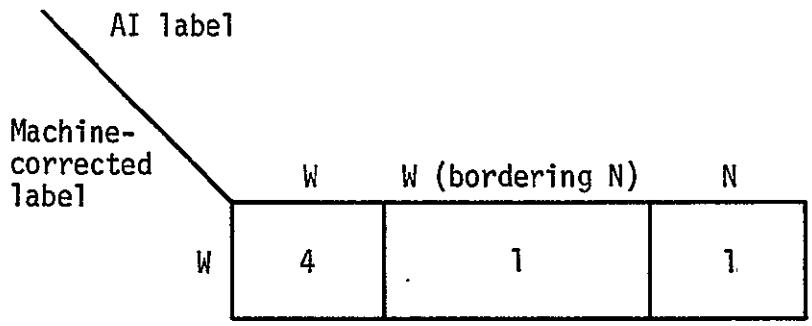

Figure 4-4.— AI labels for patterns where labels were changed from other to wheat.

To identify and change the labels of mislabeled patterns, the following scheme is proposed:

Change the label of X from $\omega' = i$ to $\omega = \ell$ if

$$g_\ell(X) = \max_{\substack{j \\ j=1,2,\cdots,M \\ j\neq i}} g_j(X) > g_i(X) + t_i \tag{4-21}$$

where t_i is a positive number.

Otherwise, do not change the label of X.

The threshold t_i for identifying the incorrect labels is determined by specifying the probability α, that mislabeling will occur in the label correction process of equation (4-21). An upper bound on the probability that such a scheme gives an incorrect label is derived as follows.

$$\begin{aligned}
P_{BL} &= P(\omega = 1)P\left[g_\ell(X) = \max_{\substack{j \\ j=1,2,\cdots,M \\ j\neq 1}} g_j(X) > g_1(X) + t_1 \,|\, \omega = 1\right] \\
&\quad + P(\omega = 2)P\left[g_\ell(X) = \max_{\substack{j \\ j=1,2,\cdots,M \\ j\neq 2}} g_j(X) > g_2(X) + t_2 \,|\, \omega = 2\right] + \cdots \\
&\quad + P(\omega = M)P\left[g_\ell(X) = \max_{\substack{j \\ j=1,2,\cdots,M \\ j\neq M}} g_j(X) > g_M(X) + t_M \,|\, \omega = M\right] \\
&= \sum_{i=1}^{M} P(\omega = i)P\left[g_\ell(X) = \max_{\substack{j \\ j=1,2,\cdots,M \\ j\neq i}} g_j(X) > g_i(X) + t_i \,|\, \omega = i\right] \\
&\leq \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} P(\omega = i)P[g_j(X) > g_i(X) + t_i \,|\, \omega = i]
\end{aligned} \tag{4-22}$$

It is assumed that the densities $p(X|\omega' = i)$ are multivariate normal. That is, $p(X|\omega' = i) \sim N(M_i', \Sigma_i')$, $i = 1,2,\cdots,M$.

Let

$$\begin{aligned} g_{ji}(X) &= g_j(X) - g_i(X) \\ &= W_j^T X + w_{j0} - W_i^T X - w_{i0} \\ &= W_{ji}^T X + w_{ji0} \end{aligned} \tag{4-23}$$

Since $g_{ji}(X)$ is a linear combination of the components of pattern vector X, if X is normally distributed, $g_{ji}(X)$ is also normally distributed. That is,

$$p[g_{ji}(X)|\omega' = s] \sim N\left[m'_{jis}, (\sigma'_{jis})^2\right] \tag{4-24}$$

where

$$\left.\begin{aligned} m'_{jis} &= W_{ji}^T M_s' + w_{ji0} \\ (\sigma'_{jis})^2 &= W_{ji}^T \Sigma_s' W_{ji} \end{aligned}\right\} \tag{4-25}$$

From equations (4-8), (4-22), and (4-25), the following is obtained.

$$\begin{aligned} P_{BL} &\leq \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} \sum_{s=1}^{M} \delta_{is} P(\omega' = s) P[g_j(X) > g_i(X) + t_i | \omega' = s] \\ &= \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} \sum_{s=1}^{M} \delta_{is} P(\omega' = s) \frac{1}{\sqrt{2\pi}\,\sigma'_{jis}} \int_{t_i}^{\infty} \exp\left[-\frac{1}{2} \frac{(y - m'_{jis})^2}{\sigma'^2_{jis}}\right] dy \\ &= \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} \sum_{s=1}^{M} \delta_{is} P(\omega' = s) \int_{-\infty}^{\frac{-t_i + m'_{jis}}{\sigma'_{jis}}} \psi(y)\,dy \end{aligned} \tag{4-26}$$

where $\psi(y)$ is given by equation (4-15). The thresholds t_i ($i = 1,2,\cdots,M$) can be determined using an optimization technique such as the Davidon-Fletcher-Powell procedure. However, it is to be noted that when $M = 2$, equations (4-17)

and (4-26) are identical. The thresholds are pictorially illustrated in figure 4-5.

Figure 4-5 shows that the imperfect label identification scheme in the multi-class case amounts to establishing a region around each decision surface.

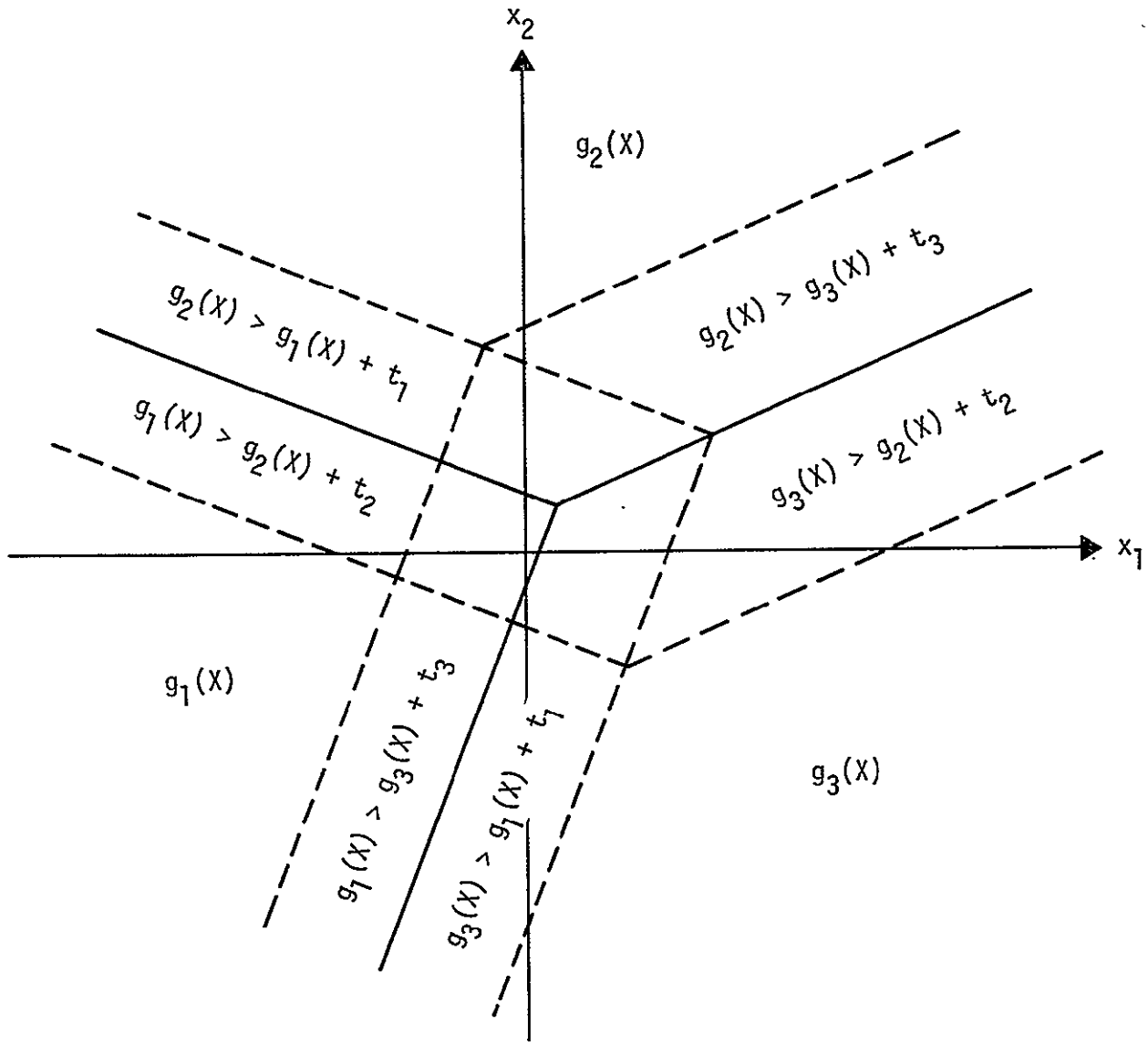

Figure 4-5.— Illustration of decision surfaces and thresholds.

5. CONCLUSIONS

In the practical applications of pattern recognition, obtaining labels for the patterns is expensive and very often these labels are imperfect. This paper has presented the problem of estimating imperfections in the labels and the use of these estimates in the identification of mislabeled patterns.

It is assumed that a set of labeled patterns, the labels of which might be imperfect, and a set of unlabeled patterns are available. The classifier classifies these patterns, and the results are a confusion matrix for the labeled pattern set and classification counts for the unlabeled set.

Expressions are presented for the maximum likelihood estimates of classification errors, for percentages of correct classification and proportions, and for the asymptotic variances of probability of correct classification and proportions.

Assuming imperfections in the labels, simple models are presented for modeling imperfections in the labels and classification errors. The problem of maximum likelihood estimation of various quantities is formulated for a general case, in terms of simplified models and class-dependent models, and their relative complexities are discussed. Results of practical applications of maximum likelihood estimation of various quantities are presented.

Assuming the densities are Gaussian and the probabilities of label imperfections are known, thresholding schemes are proposed for the identification of mislabeled patterns both for the two-class and the multiclass cases. The probability that such an identification scheme results in a wrong decision for a pattern is expressed as a function of the thresholds, and the thresholds can be computed by specifying the probability of a wrong decision by the imperfect label identification scheme.

Furthermore, the results of applying these techniques to the processing of remotely sensed multispectral data are presented.

6. REFERENCES

1. Duda, R. O.; and Singleton, R. C.: Training a Threshold Logic Unit With Imperfectly Classified Patterns. Presented at the WESCON Conv. (Los Angeles), Aug. 1964.

2. Kashyap, R. L.: Algorithms for Pattern Classification. Adaptive, Learning, and Pattern Recognition Systems, Academic Press (New York), 1970, pp. 81-113.

3. Shanmugam, K.; and Breiphol, A. M.: An Error Correcting Procedure for Learning With an Imperfect Teacher. IEEE Trans. Systems, Man and Cybernetics, July 1971, pp. 223-229.

4. Chittineni, C. B.: Learning With Imperfectly Labeled Patterns. Lockheed Electronics Co., Inc., Technical Memorandum LEC-13068, JSC-14867, NASA/JSC (Houston), Apr. 1979; Proc. 1979 IEEE Conf. on Pattern Recognition and Image Processing (Chicago) Aug. 6-9, 1979, pp. 52-62.

5. Chittineni, C. B.: On the Selection of Effective Features From the Imperfectly Labeled Patterns. Int. J. of Computer and Information Sciences, vol. 2, no. 2, 1973, pp. 103-114.

6. Chittineni, C. B.: On Feature Extraction From Imperfectly Labeled Patterns. IEEE Trans. Systems, Man and Cybernetics, May 1973, pp. 290-292.

7. Chittineni, C. B.: On the Application of Probabilistic Distance Measures for the Extraction of Features From Imperfectly Labeled Patterns. Proc. 6th Annual Princeton Conf. on Information Sciences and Systems, Mar. 1972.

8. Highleyman, W. H.: The Design and Analysis of Pattern Recognition Experiments. Bell System Tech. J., vol. 41, 1962, pp. 723-744.

9. Fukunaga, K.; and Kessell, D. L.: Nonparametric Bayes Error Estimation Using Unclassified Samples. IEEE Trans. on Information Theory, vol. IT-19, 1973, pp. 434-440.

10. Havens, K. A.; Minter, T. C.; and Thadani, S. G.: Estimation of the Probability of Error Without Ground Truth and Known a Priori Probabilities. IEEE Trans. on Geoscience Electronics, vol. GE-15, no. 3, July 1977, pp. 147-152.

11. Chow, C. K.: On Optimum Recognition Error and Reject Tradeoff. IEEE Trans. on Information Theory, vol. IT-16, Jan. 1970, pp. 41-46.

12. Chittineni, C. B.: On the Estimation of Probability of Error. Pattern Recognition, vol. 9, no. 4, 1977, pp. 191-196.

13. Chittineni, C. B.: Fisher Classifier and Its Probability of Error Estimation. LEC-13301 (JSC-14865), May 1979.

14. Rao, C. R.: Linear Statistical Inference and Its Applications. John Wiley and Sons, Inc. (New York), 1965.

15. Tenenbein, A.: Estimation From Data Subject to Measurement Error. Ph. D. dissertation, Statistics Department, Harvard University (Cambridge, Mass.), 1969.

16. Fletcher, R.; and Powell, M. J. D.: A Rapidly Convergent Descent Method for Minimization. Computer J., vol. 6, Apr. 1963, pp. 163-168.

17. Davidon, W. C.: Variable Metric Method for Minimization. Atomic Energy Commission Research and Development Report ANL-5990 (rev.), Argonne National Laboratory (Lemont, Ill.), Nov. 1959.

18. Cooper, L.; and Steinberg, D.: Introduction to Methods of Optimization. W. B. Saunders Co. (Philadelphia), 1970.