JSC-08971

ANNUAL REPORT

Statistical Theory and Methodology for Remote Sensing Data Analysis

Patrick L. Odell

Principal Investigator

June 1, 1973-June 1, 1974

THE UNIVERSITY OF TEXAS AT DALLAS

Dallas, Texas

STATISTICAL THEORY AND METHODOLOGY FOR REMOTE SENSING DATA ANALYSIS

ANNUAL REPORT

ACKNOWLEDGMENTS

The research work documented in this final report is carried out for NASA Johnson Space Center, Houston, Texas, under Contract NAS9-13512 to The University of Texas at Dallas, Richardson, Texas, for the period June 1, 1973 to May 31, 1974. This was accomplished in collaboration with Dr. J. P. Basu and Dr. R. S. Chhikara, research scientists, and Mr. Lynn Ziegler, graduate student, all of UTD; and Dr. T. L. Boullion of the Texas Tech University Statistics Faculty and his graduate student, Mr. E. R. Knezek, Jr.

Patrick L. Odell
Principal Investigator

PREFACE

With the recent developments in statistical methodology of clustering and classification as well as in automatic remote sensing data processing techniques, the scope of extracting useful information on earth resources from multispectral scanner data has increased considerably. Though the performance of most of the previously developed remote sensing data handling techniques has yet to be ascertained through the process of testing and numerical evaluations, it now appears that the development of a technology for performing a large area earth resources inventory is in sight.

For the earth resources project of EOD, NASA-JSC, presently a top priority has been given to the development of a system for performing an inventory of some crop or crops of economic interest over a large area. An important aspect of this project would be to devise a suitable crop acreage estimation procedure. A major part of the research work reported here in our final report is concerned with this problem.

First of all, a model is developed for the evaluation of crop acreages (proportions) in an agricultural area using the classification approach. The model takes into account the classification errors likely to arise in labeling remotely sensed data points under the classification algorithm used and evaluates the actual crop proportions by correcting the expected labeled crop proportions for the possible bias due to these errors.

If the goal is to determine crop acreages for a large area, it will be necessary to use only a subset of the full data obtained using a sampling technique since it would not be practical to collect and process a complete set of scanner data covering the entire area. Depending upon whether or not

any information is available on the area crop layouts, agricultural practices, etc., a suitable sampling scheme needs to be devised for acquiring a well representative sample of the unlabeled remotely sensed data. The precision of any crop acreage estimates will also depend upon the performance of the classification algorithm used in labeling remotely sensed data points and whether or not the associated classification errors are known. It is very unlikely to have these errors known in advance. As such, a certain amount of ground truth, preferably a sample of ground truth for each crop type in the area of interest, needs to be ascertained so that the classifier is properly trained and the classification errors estimated.

Considering these aspects of the problem, we have discussed the estimation of crop acreages for different possible cases. If the classification errors for the classification procedure used are assumed known, the estimation method provides best estimates for the crop acreages. In case of unknown classification errors, the estimates are consistent. Next, both the error analysis and the problem of sample size are investigated in general as well as for certain specific cases. Our results are given in reports 1, 4 and 5 listed in the table of contents.

A study of classification errors for the Gaussian maximum likelihood classifier is made when due to interest in identifying elements of only one class, the other class is made of the remainder of the classes and then the problem is treated as a two-class classification problem. Given in report 6, depending upon the geometry and location of the classes under this practice, it is shown that the classification performance for elements in the class of main interest may or may not improve under this practice.

The estimation of optimum errors of classification and the dependence of these estimates upon the distance between classes are examined for the two-class problem, assuming classes to be univariate normal, in report 2. Considering both the maximum likelihood estimate and the minimum variance unbiased estimate for the optimum probability of misclassification, we give the relative efficiencies of these two estimates, theoretically as well as numerically, and investigate the bias of the former.

A simulation study is done on the relationship between the probability of misclassification using linear as well as quadratic discriminant rules, the number of features and the training sample size. As shown in report 3, when the sample size is small, use of a fewer number of features for discrimination leads to smaller probability of misclassification.

For classifying an observation into one of two given normal populations whose parameters are unknown, the usual practice in the absence of any training sample is to cluster past data into two nearest neighbor clusters and to design a sample based Bayes' classifier, treating the two clusters as training samples from the two populations. The use of ℓ_1-norm is often advocated for such clustering. In report 7 it has been shown that such advocation is not always reasonable.

A unified theory of adaptive pattern recognition has been presented in report 8. It has been shown that all adaptive pattern recognition algorithms are just different means of approximating a function that separates the sets of training samples, choosing different criteria for best approximation.

The data obtained by remote sensing devices in Earth Resources Survey come from a large area and often over a large period of time. Due to changes

in spatial and temporal conditions the statistical characteristics of the data have been found to undergo changes. In such a situation, the performance of sample-based Bayes classifier designed on the assumption of normality of the populations can be enhanced by periodic updating of the parameter estimates. It has been shown in report 9 that all updating algorithms that can be found in literature are related to one particular model of the random environment.

TABLE OF CONTENTS

ACKNOWLEDGMENTS i
PREFACE ii

REPORTS

1. ESTIMATION OF A LARGE AREA CROP ACREAGE INVENTORY USING REMOTE SENSING TECHNOLOGY (See #10.) 1
R. S. Chhikara and P. L. Odell

2. ESTIMATION OF OPTIMUM ERRORS OF CLASSIFICATION FOR UNIVARIATE NORMAL POPULATIONS 49
R. S. Chhikara and P. L. Odell

3. A SIMULATION STUDY OF POPULATION CLASSIFICATION USING SELECTED VARIABLES 66
E. R. Knezek, Jr. and T. L. Boullion

4. ACREAGE ESTIMATES FOR CROPS USING REMOTE SENSING TECHNIQUES: CLASSIFICATION ERROR MATRIX KNOWN (See #10) 93
R. S. Chhikara and P. L. Odell

5. ESTIMATION OF CROP ACREAGE THROUGH SAMPLING OF REMOTELY SENSED DATA: CLASSIFICATION ERROR MATRIX UNKNOWN (See #10.) 106
R. S. Chhikara and P. L. Odell

6. ON COMBINING POPULATIONS IN STATISTICAL CLASSIFICATION USING REMOTE SENSING DATA 123
J. P. Basu and P. L. Odell

7. EFFECT OF DISTANCE MEASURE ON CLUSTER BASED CLASSIFICATION PROCEDURES 138
J. P. Basu and P. L. Odell

8. ADAPTIVE PATTERN RECOGNITION--A SURVEY 143
J. P. Basu and P. L. Odell

9. ON RECOGNITION OF WANDERING PATTERNS 162
J. P. Basu and P. L. Odell

10. ADDENDUM TO PAPERS 1, 4 AND 5 186

ESTIMATION OF A LARGE AREA CROP ACREAGE INVENTORY USING REMOTE SENSING TECHNOLOGY*

R. S. Chhikara and P. L. Odell

The University of Texas at Dallas

* This research is carried out for NASA and supported by Contract NAS9-13512

ESTIMATION OF A LARGE AREA CROP ACREAGE INVENTORY USING REMOTE SENSING TECHNOLOGY

0. SUMMARY

Based upon the existing remote sensing capabilities, the useful information about the acreage of some crop of economic interest can be obtained from multispectral scanner measurements acquired over an agricultural area. If the goal is to determine the acreages covered by various crops over some large area such as the continental United States, then some sampling procedure will be necessary since it would not be practical to collect and process a set of scanner data covering the entire area.

In this report we develop a model for the evaluation of acreages (proportions) of different crop-types over a geographical area using a classification approach and give methods for estimating the crop acreages. If prior information is available on the classification errors associated with the classification algorithm used, the estimation method provides the best estimate for the crop acreages. Otherwise, the method would first require a certain amount of ground truth in the area of interest to be obtained so that the classifier can be trained and the classification errors estimated.

If the main interest lies in estimating the acreages of a specific crop-type such as wheat, it is suggested to treat the problem as a two-crop problem: wheat vs. non-wheat, since this simplifies the estimation problem considerably. The error analysis and the sample size problem is investigated for the two-crop approach. Certain numerical results for sample sizes are given for a JSC-ERTS-1 data example on wheat identification performance in Hill County, Montana and Burke County, North Dakota. Lastly, for a large area crop acreages inventory we suggest a sampling scheme for acquiring sample data and discuss the problem of crop acreage estimation and the error analysis.

1. INTRODUCTION

In recent years the development of several automatic data processing techniques for statistical pattern recognition has enhanced considerably the scope of remote sensing technology for the study of earth resources, particularly in the field of agriculture. It now appears that a system for performing a large area crop inventory can be developed on the basis of existing remote sensing capabilities.

The data handling and analysis for remotely sensed agricultural resources over a large area may not be feasible both from technical and economical viewpoints if each scanned data point is being processed for its recognition. For example, if a complete recognition is desired for an ERTS scene, it would require processing over half a million data points. As such, an important requirement for any system to be developed for a large area crop inventory should be to have a suitable crop acreage estimation technique that uses only a sample of the unlabeled remotely sensed data obtained for the area of interest for the purpose of recognition.

In this report we discuss a large area crop acreage estimation procedure that would meet this requirement for the system. We develop a model for the evaluation of crop proportions for an agricultural area and provide methods for crop acreage estimation, taking into consideration the classification errors likely to arise in labeling remotely sensed data. The error analysis for the model is studied and expressions for variances of different estimates are given, in general as well as in specific cases. For the two-crop situation, the problem of sample size is investigated and certain numerical results for the sample size are provided. Next, we extend the scope of our study to investigate a large area crop inventory.

2. CROP PROPORTIONS MODEL

Suppose there are m different crops $\pi_1, \pi_2, \ldots, \pi_m$ in the agricultural area of interest and that every data point is identifiable with respect to one of these crops. Let p_i denote the proportion of pixels in π_i, i=1,2,...,m. Considering a random sample of n unlabeled remotely sensed data points, let n_i be the number of points classified into π_i, i=1,2,...,m, using a classification algorithm. Suppose $n(i|j)$ is the number of data points to be in π_j but classified into π_i, then

$$n_i = n(i|1) + n(i|2) + \ldots\ldots + n(i|m)$$

and

$$\frac{n_i}{n} = \sum_{j=1}^{m} \frac{n(i|j)}{n} \quad , \qquad i=1,2,\ldots,m \tag{2.1}$$

are the observed crop proportions for the sample data under the classification algorithm used. The observed proportion n_i/n is a biased estimate of p_i since it estimates unbiasedly $E[n_i/n]$ given by

$$e_i = \sum_{j=1}^{m} E\left[\frac{n(i|j)}{n}\right]$$

$$= \sum_{j=1}^{m} p_j \, P(i|j) \tag{2.2}$$

where $P(i|j)$ denotes the probability of classifying a data point from π_j into π_i under the classification algorithm. It may be pointed out that processing of remotely sensed data for total recognition would lead to an evaluation of the expected classified crop proportions e_i's instead of the

actual crop proportions p_i's. Of course, if the classification algorithm performs so well that the classifiction errors are sufficiently small, e_i will be close enough to p_i, i=1,2,...,m. But most statistical pattern recognition techniques for processing of remotely sensed data are expected to be fallible and thereby the two types of proportions are not going to be near equal. Henceforth in our discussion we will assume that $P(i|j) > 0$ for at least one j different from i.

Denoting the observed proportion n_i/n by $\hat{e}_i$,i=1,2,...,m, it follows from (2.2) that

$$e = E[\hat{e}]$$

or

$$e = Pp \tag{2.3}$$

where

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix}, \qquad p = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{bmatrix}$$

and

$$P = \begin{bmatrix} P(1|1) & P(1|2) & \cdots & P(1|m) \\ P(2|1) & P(2|2) & \cdots & P(2|m) \\ \cdots & \cdots & \cdots & \cdots \\ P(m|1) & P(m|2) & \cdots & P(m|m) \end{bmatrix} .$$

Accordingly, the vector of actual crop proportions

$$p = P^{-1}e \tag{2.4}$$

are obtained subject to $\sum_{i=1}^{m} p_i = 1$ provided e and P are known.

The vector of classified crop proportions, e, can only be known if the complete set of remotely sensed data is processed for total recognition using a classification algorithm. Hence, in general, e will be unknown. Regarding the classification error matrix P, two cases arise:

(i) P is known

(ii) P is unknown .

If P is known, an unbiased estimate of p is

$$\hat{p} = P^{-1}\hat{e} \ . \tag{2.5}$$

But P will generally be unknown. As such it would be necessary to obtain a certain amount of suitably selected gound truth in the area of interest, probably independent of the sample data used in estimating e, so that the classifier is trained and the classification error matrix estimated. Let $\hat{P}$ be an estimate of P. Then an estimate of p when P is unknown is given by

$$\hat{\hat{p}} = \hat{P}^{-1}\hat{e} \ . \tag{2.6}$$

Clearly, $\hat{\hat{p}}$ will generally be a biased estimate. Both bias and mean square error of $\hat{\hat{p}}$ will depend upon the performance of the classification algorithm as well as the degree to which the sample represents the population. The classification performance can be achieved desiredly by training the classifier sufficiently on the basis of a well representative sample for the ground truth. By adopting certain sampling schemes that may be suitable for the area of interest, appropriate samples for both the ground truth and the unlabeled remotely sensed data can be acquired. For our later discussion we assume that these two types of samples are independently obtained.

In appendix 1 we have derived general expressions for the covariance matrix of $\hat{p}$, and both the bias and the mean square error matrix of $\hat{\hat{p}}$. In the

case of $\hat{\hat{p}}$ not only the estimate $\hat{\hat{p}}$ itself but bias as well as mean square error quantities will also depend upon how $\hat{P}$ is obtained. One solution for $\hat{P}$ and the probability distribution of its components is suggested therein, as well.

3. TWO-CROP APPROACH

Sometimes the main interest is in estimating the acreage of a specific crop type in the area of interest. In that case one approach to the acreage estimation problem lies in considering π_1 to be the specific crop type and π_0 to be the "other crop" consisting of the remainder of the crops, and then treating it as a two-crop situation. However, lumping of different crops together for the "other crop" would require certain caution and should be judged in terms of the classification performance for the two-crop case as against that for the case of the original set of crops. For the Gaussian maximum likelihood classifier, Basu and Odell (1973) have investigated this problem and have shown that the classification performance for the class of main interest may or may not improve when the classification is performed using the two-class approach. But the problem under this approach is greatly simplified and, barring extreme cases, perhaps it will provide satisfactory solutions in the remote sensing situation when interest only lies in ascertaining the acreage cover of one specific crop.

Now considering two crops π_1 and π_0, let $P(1|0) = \Phi_1$ and $P(0|1) = \Phi_2$ for the probabilities of misclassification when a certain classifier is used. We will assume that $\Phi_1 + \Phi_2 \neq 1$. Then

$$P = \begin{bmatrix} 1 - \Phi_2 & \Phi_1 \\ \Phi_2 & 1 - \Phi_1 \end{bmatrix} .$$

If p_1 and p_2 are the actual crop proportions of π_1 and π_0, respectively, whereas e_1 and e_2 are their respective expected classified crop proportions under the classifier used, it follows from (2.3) that

$$e_1 = (1-\Phi_2)\ p_1 + \Phi_1\ (1-p_1) \tag{3.1}$$

and

$$e_2 = 1-e_1 \ .$$

On the other hand,

$$p_1 = \frac{e_1 - \Phi_1}{1-\Phi_1-\Phi_2} \tag{3.2}$$

and

$$p_2 = \frac{e_2 - \Phi_2}{1-\Phi_1-\Phi_2} \qquad \text{or} \quad p_2 = 1-p_1.$$

Suppose from a random sample of n unlabeled remotely sensed data points, n_1 points were classified into π_1 and $n_2 = n-n_1$ points were classified into π_0 by the classifier. Then

$$\hat{e}_1 = \frac{n_1}{n}\ , \tag{3.3}$$

$$\hat{p}_1 = \frac{\hat{e}_1 - \Phi_1}{1-\Phi_1-\Phi_2} \tag{3.4}$$

if Φ_1 and Φ_2 are known, and

$$\hat{\hat{p}}_1 = \frac{\hat{e}_1 - \hat{\Phi}_1}{1-\hat{\Phi}_1-\hat{\Phi}_2} \tag{3.5}$$

when Φ_1 and Φ_2 are unknown and have estimates $\hat{\Phi}_1$ and $\hat{\Phi}_2$ respectively. Clearly

$$\mathrm{Var}(\hat{p}_1) = \frac{1}{(1-\Phi_1-\Phi_2)^2}\ \mathrm{Var}(\hat{e}_1)\ . \tag{3.6}$$

For the estimate $\hat{\hat{p}}_1$ in (3.5), it easily follows that

$$\text{Bias } (\hat{\hat{p}}_1) = (e_1-\Phi_1)\, E[T-\theta] - E[T(\hat{\Phi}_1-\Phi_1)] \tag{3.7}$$

where

$$T = (1-\hat{\Phi}_1-\hat{\Phi}_2)^{-1}$$

and

$$\theta = (1-\Phi_1-\Phi_2)^{-1} .$$

However, evalution of expected values in (3.7) may be quite a difficult job and so an exact bias value may not be accessible. An evaluation of the $MSE(\hat{\hat{p}}_1)$ in its exact form is even more difficult. As such we instead consider having it in the following approximate form obtained in Appendix 1 using the δ-method. For a discussion on the method, see Rao (1965).

$$MSE(\hat{\hat{p}}_1) = \frac{1}{[1-\Phi_1-\Phi_2]^2} \left(Var(\hat{e}_1) + \left[1 - \frac{e_1 - \Phi_1}{1-\Phi_1-\Phi_2}\right]^2 Var(\hat{\Phi}_1) + \left[\frac{e_1 - \Phi_1}{1-\Phi_1-\Phi_2}\right]^2 Var(\hat{\Phi}_2)\right) \tag{3.8}$$

or

$$MSE(\hat{\hat{p}}_1) = \frac{1}{[1-\Phi_1-\Phi_2]^2} \left[Var(\hat{e}_1) + (1-p_1)^2 Var(\hat{\Phi}_1) + p_1^2\, Var(\hat{\Phi}_2)\right] \tag{3.9}$$

where p_1 is given by (3.2).

Sample Size

Considering simple random sampling with pixel as the sampling unit, we discuss the problem of sample size necessary to minimize the sampling cost or to achieve a desired amount of precision for the proportion estimate, given that the other is specified. Suppose total sample consists of $N = n + N_1 + N_2$ data points selected randomly, where n is the size of sample of unlabeled remotely sensed data used for estimating e_1, and N_1 and N_2 are sample sizes for ground truth data from π_1 and π_0 and are used to estimate Φ_1 and Φ_2, respectively. The estimates $\hat{e}_1$, $\hat{\Phi}_1$ and $\hat{\Phi}_2$ are all obtained as observed sample proportions and thus it follows from (3.6) and (3.9) that

$$\mathrm{Var}(\hat{p}_1) = \frac{e_1(1 - e_1)}{n(1-\Phi_1-\Phi_2)^2}$$

and

$$\mathrm{MSE}(\hat{\hat{p}}_1) = \frac{1}{(1-\Phi_1-\Phi_2)^2}\left[\frac{e_1(1-e_1)}{n} + (1-p_1)^2\frac{\Phi_1(1-\Phi_1)}{N_1} + p_1^2\frac{\Phi_2(1+\Phi_2)}{N_2}\right] . \qquad (3.10)$$

Suppose we want to obtain sample sizes necessary to minimize the sampling cost when $\mathrm{Var}(\hat{p}_1)$ and $\mathrm{MSE}(\hat{\hat{p}}_1)$ are specified, say each equal to or smaller than σ^2. In the case of Φ_1, Φ_2 known, the only cost involved is that of processing the remotely sensed sample data. Clearly, it will be minimum when the sample size n is the smallest integer greater than or equal to

$$\frac{e_1(1-e_1)}{(1-\Phi_2-\Phi_2)^2\,\sigma^2} . \qquad (3.11)$$

For when Φ_1 and Φ_2 are unknown, there are two types of cost involved: one is the cost of processing the total sample data, say at the rate of c_1 dollars per data point and the other is the cost of obtaining ground truth, say at the rate of c_2 dollars per data point. Then the cost associated with a sample of size $N = n + N_1 + N_2$ is of the form:

$$C(N) = c_1 n + (c_1+c_2)\,(N_1+N_2) . \qquad (3.12$$

The purpose is to find N (i.e., n, N_1 and N_2) which minimize C(N) subject to $\mathrm{MSE}(\hat{\hat{p}}_1) \leq \sigma^2$. This is done in Appendix 2 where we derive explicit expressions for n, N_1 and N_2 in (A.9).

4. AN EXAMPLE

Certain sites in Hill County, Montana and Burke County, North Dakota were selected to investigate wheat identification performance for the ERTS-1 satellite data during 1973. For the sites in Hill County, there were three acquisition periods, covering both winter and spring wheat seasons, for which ERTS-1 labeled data were evaluated against the ground truth to ascertain wheat identification performance. In the case of the site in Burke County, there were only two acquisition periods covering the spring wheat season. For the classification identification performance results and other details, refer to Appendix 3.

Considering Φ_1 to be the omission percentage for the non-wheat data points and Φ_2 for the wheat data points, we give sample size results in Figure 1-7 for the various cases of omission percentages listed in Appendix 3, assuming different wheat proportions in the area and $\sigma_2 = .01$. Based on these results, the following conclusions are drawn:

1. Expected labeled wheat proportion, e_1, increases as the actual proportion of wheat, p_1, increases for the area, though not strictly. Though to a certain extent it depends upon the magnitude of the omission percentages for both non-wheat and wheat data points, it tends to centralize away from too low or too high values for the percentage.
2. Sample size for the unlabeled remotely sensed data first increases as the actual wheat proportion increases and then decreases later on; the point of decrease depends upon the size of the two omission percentages.

3. All sample sizes increase as the total omission rate $\Phi_1 + \Phi_2$ increases.
4. Sample size for the unlabeled remotely sensed data is much larger when Φ_1, Φ_2 are unknown compared to when these are known.
5. In the case of Φ_1, Φ_2 unknown, the sample size for the unlabeled remotely sensed data is proportional to c_2/c_1, the ratio of two types of cost.
6. Sample sizes for ground truth of wheat and non-wheat are inversely proportioned to c_2/c_1.
7. Sample size for the ground truth of wheat is larger than that for non-wheat when the expected labeled wheat proportion is below .5. Reverse is the case when such proportion is above .5. A similar trend holds against the actual wheat proportion, though not strictly.
8. Sample size for the ground truth increases as either of the two omission percentages increases when the other is held fixed.

For making a comparison of sample sizes irrespective of the wheat proportion which, in fact, is unknown, a suitable criterian is to determine the sample sizes against values for the coefficient of variation, C.V. = σ/p. Generally the wheat coverage in any area of interest is expected to be somewhere in between 1 percent and 20 percent. As such we here give sample sizes for the unlabeled remotely sensed data and the ground truth of wheat as well as non-wheat by specifying σ = .01 and considering certain C.V. values in a 5 to 50 percent range. Numerical results are presented in Table 1 for all different cases of Φ_1, Φ_2 values that arise from the wheat identification performance

results given in Appendix 3. Moreover, for certain cases the sample sizes are sketched in Figure 8-14. The following conclusions are drawn:

1. All samples sizes increase as the total omission percentage $\Phi_1 + \Phi_2$ increases.
2. Except for the sample size for the ground truth of wheat, sample sizes decrease as the coefficient of variation increases. These are generally very high in numbers for the 5 percent co-efficient of variation but levels off when the co-efficient of variation is 50 percent.
3. Sample size for the unlabeled remotely sensed data increases considerably if Φ_1, Φ_2 are unknown compared to their known case.
4. Again, all sample sizes depend upon the ratio c_2/c_1 as regards the two types of cost.
5. Sample size for the ground truth of wheat is consistently larger than that of non-wheat. Also, it shows very small changes over the range of co-efficients of variations being considered here. In cases where there is a high overall omission percentage, and particularly for the non-wheat, it tends to increase as the co-efficient of variation increases.

TABLE 1: Sample sizes: n for the unlabeled remotely sensed data, N_1 for the ground truth of wheat, and N_2 for the ground truth of non-wheat when $\sigma = .01$

Coefficient of variation	Wheat proportion P_1	Omission rates Φ_1	Omission rates Φ_2	Expected labeled wheat proportion e_1	Φ_1, Φ_2 known Sample size n	Φ_1 and Φ_2 unknown case: $c_2/c_1=5$ Sample size n	$c_2/c_1=5$ Ground truth sample sizes N_1	$c_2/c_1=5$ Ground truth sample sizes N_2	$c_2/c_1=20$ Sample size n	$c_2/c_1=20$ Ground truth sample sizes N_1	$c_2/c_1=20$ Ground truth sample sizes N_2
0.050	0.200	0.200	0.300	0.3000	8400	26884	7664	2195	42979	550	876
		0.100	0.250	0.2300	4192	12161	2832	1022	19100	2377	56
		0.150	0.100	0.3000	3734	10832	2700	569	16638	2264	476
		0.100	0.150	0.2500	3334	9206	2083	620	14319	1732	16
		0.050	0.200	0.2000	2645	7275	1295	598	11134	1059	466
		0.050	0.100	0.2200	2376	5667	974	336	8533	784	70
		0.000	0.050	0.1900	1706	2170	0	99	2574	0	3
0.100	0.100	0.200	0.300	0.2500	7500	24718	8390	1068	39712	7205	918
		0.100	0.250	0.1650	3201	11004	2971	477	15875	2520	405
		0.150	0.100	0.2250	3100	9453	2582	279	15054	2523	230
		0.100	0.150	0.1750	2567	7625	2212	293	12030	1806	247
		0.050	0.200	0.1250	1545	5346	1295	264	8308	1076	220
		0.050	0.100	0.1350	1617	4237	993	152	6518	817	125
		0.000	0.050	0.0950	953	1127	0	35	1278	0	21
0.150	0.067	0.200	0.300	0.2333	7156	23893	8610	705	38469	7410	607
		0.100	0.250	0.1433	2937	9162	2995	309	14646	2554	264
		0.150	0.100	0.2000	2845	8998	3061	184	14357	2611	157
		0.100	0.150	0.1500	2267	6991	2236	191	11105	1901	162
		0.050	0.200	0.1000	1600	4006	1275	160	7224	1069	141
		0.050	0.100	0.1067	1319	3658	984	97	5694	819	81
		0.000	0.050	0.0633	658	754	0	19	838	0	11
0.200	0.050	0.200	0.300	0.2250	6975	23460	8716	526	37816	7510	453
		0.100	0.250	0.1325	2721	8749	3003	229	13997	2568	186
		0.150	0.100	0.1875	2705	8729	3098	137	13972	2650	118
		0.100	0.150	0.1375	2109	6651	2247	141	10606	1916	120
		0.050	0.200	0.0875	1420	4214	1261	122	6647	1063	103
		0.050	0.100	0.0925	1162	3343	976	71	5243	818	60
		0.000	0.050	0.0475	502	565	0	12	628	0	7
0.250	0.040	0.200	0.300	0.2200	6864	23194	8778	419	37414	7569	362
		0.100	0.250	0.1260	2607	8481	3005	181	13597	2575	155
		0.150	0.100	0.1800	2624	8560	3118	110	13729	2673	94
		0.100	0.150	0.1300	2011	6438	2251	112	10293	1924	96
		0.050	0.200	0.0800	1309	3970	1250	96	6267	1058	81
		0.050	0.100	0.0840	1065	3146	969	56	4958	816	47
		0.000	0.050	0.0380	406	451	0	9	490	0	5
0.300	0.033	0.200	0.300	0.2167	6789	23014	8619	349	37142	7608	301
		0.100	0.250	0.1217	2530	8300	3006	150	13324	2560	129
		0.150	0.100	0.1750	2567	8444	3132	91	13561	2669	78
		0.100	0.150	0.1250	1945	6253	2253	93	10079	1929	80
		0.050	0.200	0.0750	1234	3803	1242	79	6041	1055	67
		0.050	0.100	0.0783	1000	3010	964	46	4761	815	39
		0.000	0.050	0.0317	340	375	0	7	405	0	4
0.350	0.029	0.200	0.300	0.2143	6735	22824	8847	299	36946	7635	258
		0.100	0.250	0.1186	2474	8168	3006	128	13127	2583	110
		0.150	0.100	0.1714	2520	8359	3141	78	13439	2700	67
		0.100	0.150	0.1214	1897	6187	2254	79	9923	1933	68
		0.050	0.200	0.0714	1180	3682	1230	67	5862	1052	57
		0.050	0.100	0.0743	952	2911	960	39	4616	814	33
		0.000	0.050	0.0271	293	321	0	6	344	0	3
0.400	0.025	0.200	0.300	0.2125	6694	22785	8868	261	36797	7656	225
		0.100	0.250	0.1163	2432	8069	3006	112	12977	2565	96
		0.150	0.100	0.1688	2494	8295	3140	68	13346	2708	59
		0.100	0.150	0.1138	1861	6107	2255	69	9805	1935	60
		0.050	0.200	0.0687	1139	3590	1231	58	5725	1050	50
		0.050	0.100	0.0712	916	2835	956	34	4506	813	29
		0.000	0.050	0.0237	257	280	0	5	300	0	3
0.450	0.022	0.200	0.300	0.2111	6662	22708	8885	232	36681	7672	200
		0.100	0.250	0.1144	2389	7991	3006	95	12860	2586	85
		0.150	0.100	0.1667	2470	8244	3153	61	13272	2714	52
		0.100	0.150	0.1167	1833	6044	2255	61	9712	1937	53
		0.050	0.200	0.0667	1107	3516	1227	52	5618	1040	44
		0.050	0.100	0.0689	888	2775	954	30	4419	812	26
		0.000	0.050	0.0211	229	248	0	4	265	0	2
0.500	0.020	0.200	0.300	0.2100	6636	22646	8898	209	36583	7684	180
		0.100	0.250	0.1130	2373	7928	3006	89	12766	2587	77
		0.150	0.100	0.1650	2450	8203	3157	55	13213	2719	47
		0.100	0.150	0.1150	1810	5993	2255	55	9637	1938	48
		0.050	0.200	0.0650	1081	3460	1224	46	5531	1046	40
		0.050	0.100	0.0670	866	2727	951	27	4348	811	23
		0.000	0.050	0.0190	207	223	0	3	237	0	2

5. A LARGE AREA CROP ACREAGE ESTIMATION

Our previous discussion, in essence, applies to crop acreage inventory for an agricultural area which is homogeneous in respect to agricultural practices and thus is not expected to be large enough. Since a major objective of the JSC-EOD project is to perform or estimate crop acreages for a large area using available remote sensing capabilities, we here suggest a sampling procedure to procure sample data for the purpose of estimating a large area crop acreage inventory and discuss the error analysis associated with it.

Once again, we assume that the frame is made of agricultural areas; the non-agricultural areas in the region of interest can be easily excluded by way of a monitoring system. As a first step in the sampling procedure, we suggest having a geographical-based stratification which effects a division of the region into reasonably homogeneous areas with respect to physical and climatological conditions. Considering additional factors of (i) the predominance of various crop-types and (ii) the latitude and longitude, a finer stratification must be achieved. This is to obtain better discrimination for the underlying crop-types and to control variability which may otherwise dominate over the distinction that exists between the resolution classes for these crop-types.

Note that as a result of stratification one may only need to consider a part of the region for frame if crops of interest do not cover the whole region. So depending upon whether the frame would require consideration of the complete region or only a part of it, one should make a list of strata making up the frame for the purpose of sampling.

Remoting sensing data gathered by an ERTS satellite is documented in terms of scenes, each covering approximately an area of 100 × 100 miles and

divided into four strips where each strip has approximately 6,400 scanlines in it. As such, we suggest a three stage sampling plan to be independently carried out in each stratum: select randomly ERTS scenes at the first stage, strips within scenes at the second stage and scanlines within strips at the third stage. Of course, one may consider one more stage in selecting pixels within scanlines. However, sampling at this stage is excluded from the plan because it is inconvenient and uneconomical.

Notations

Let R be the region (in the sense of frame) of interest for estimating crop acreages. Suppose it is stratified into strata R_t, t=1,2,...,L, with weights w_t, the proportion of pixels in tth stratum, t=1,2,...,L so that

$$R = \bigcup_{t=1}^{L} R_t \qquad \text{with} \quad \sum_{t=1}^{L} w_t = 1 \quad .$$

In stratum R_t, let I_t be the number of scenes whereas J, H and n denote the number of strips per scene, number of scanlines per strip and number of pixels per scanlines, respectively. From the previous paragraph it is obvious that there is no need to distinguish between strata in the categories of strips per scene, scanlines per strip and pixels per scanline. Next, let $e_{tijh}(\pi_k)$ be the expected proportion of pixels to be classified in π_k from the hth scanline in jth strip of ith scene for stratum R_t, t=1,2,...,L. Then for R_t,

$$e_{tij}(\pi_k) = \sum_{h=1}^{H} e_{tijh}(\pi_k)$$

the expected proportion of pixels to be classified in π_k from the jth strip in ith scene,

$$e_{ti}(\pi_k) = \sum_{j=1}^{J} \sum_{h=1}^{H} e_{tijh}(\pi_k),$$

the expected proportions of pixels to be classified in π_k from the ith scene,

$$e_t(\pi_k) = \sum_{i=1}^{I_t} \sum_{j=1}^{J} \sum_{h=1}^{H} e_{tijh}(\pi_k) ,$$

the expected proportion of pixels to be classified in π_k. Accordingly,

$$e(\pi_k) = \sum_{t=1}^{L} w_t e_t(\pi_k) , \tag{5.1}$$

is the expected proportion of pixels to be classified in π_k, k=1,2,...,m, for the region R.

Estimates

Suppose m_t, r and s denote the corresponding number of scenes, number of strips per scenes and number of scanlines per strip that one selected for R_t, t=1,2,...,L, using the stratified three stage random sampling described earlier. Let $n_{tijh}(\pi_k)$ be the number of pixels classified into π_k from the hth selected scanline in jth selected strip of the ith selected scene in R_t. Then considering the observed proportions of classified data points into different crops for estimates, one has

$$\hat{e}_{tijh}(\pi_k) = \frac{n_{tijh}(\pi_k)}{n} ,$$

$$\hat{e}_{tij}(\pi_k) = \frac{1}{ns} \sum_{h=1}^{s} n_{tijh}(\pi_k),$$

$$\hat{e}_{ti}(\pi_k) = \frac{1}{nsr} \sum_{j=1}^{r} \sum_{h=1}^{s} n_{tijh}(\pi_k) ,$$

$$\hat{e}_{t}(\pi_k) = \frac{1}{nsrm_t} \sum_{i=1}^{m_t} \sum_{j=1}^{r} \sum_{h=1}^{s} n_{tijh}(\pi_k) ,$$

and

$$\hat{e}(\pi_k) = \sum_{t=1}^{L} w_t \hat{e}_t(\pi_k) , \qquad k = 1,2,\ldots,m . \tag{5.2}$$

Next, expressions for $\mathrm{Var}(e_t(\pi_k))$ and $\mathrm{Cov}(e_t(\pi_k), e_t(\pi_k))(k \neq k')$ can be obtained without much difficulty. For example, refer to Section 10.8 in Cochran (1963) for the general discussion on three stage sampling plan. Hence, the covariance matrix of $\hat{e}$ is given by

$$\mathrm{Var}(\hat{e}(\pi_k)) = \sum_{t=1}^{L} w_t^2 \, \mathrm{Var}(\hat{e}_t(\pi_k))$$

and

$$\mathrm{Cov}(\hat{e}(\pi_k), \hat{e}(\pi_{k'})) = \sum_{t=1}^{L} w_t^2 \, \mathrm{Cov}(\hat{e}_t(\pi_k), \hat{e}_t(\pi_{k'})), \; k \neq k', \quad k=1,2,\ldots,m . \tag{5.3}$$

Similarly, an estimate of the covariance matrix is obtained by replacing the unknown quantities by their estimates in (5.3). In this context, see Chhikara and Odell (1974) who have discussed such results in greater details.

Now to obtain the actual crop proportions, there is a need to consider whether or not the classification error matrix is the same for each stratum. When the area is wide and large and the stratification is performed considering factors mentioned in the beginning of this section, it is quite likely that these classification error matrices will not be the same for different strata.

In that case, find an estimate of p_k, k=1,2,...,m, using (2.5) if the classification error matrix is known and (2.6) if it is unknown for each stratum. Denoting p_k by $p_t(\pi_k)$ for stratum R_t, it then follows that

$$\hat{p}_k = \sum_{t=1}^{L} w_t \hat{p}_t(\pi_k) \quad , \qquad k = 1,2,\ldots,m \tag{5.4}$$

when the classification matrices, say P_t, t = 1,2,...,L, are known, and

$$\hat{\hat{p}}_k = \sum_{t=1}^{L} w_t \hat{\hat{p}}_t(\pi_k) \quad , \qquad k = 1,2,\ldots,m \tag{5.5}$$

when these are unknown and are separately estimated using ground truth data from each stratum. Next, $\mathrm{Var}(\hat{p}_k)$ and $\mathrm{MSE}(\hat{\hat{p}}_k)$ are respectively obtained from (A.2) and (A.7) after making an appropriate substitution from (5.2).

On the other hand, either there is the same classification error matrix for all strata or can be made so by proper adjustment of signatures in the classification algorithm for each stratum. For then an estimate of crop proportions p_k, k=1,2,...,m is directly given by (2.5) if the common classification error matrix is known and by (2.6) if it is unknown, using $\hat{e}(\pi_k)$, k=1,2,...,m of (5.1) for e. Hence, both estimates and their error analyses are obtained by following the general procedure given in Section 2.

In fact, our approach in Section 2 is quite general and can be applied to perform any large area acreage inventory by considering an appropriate sampling scheme for both the unlabeled remotely sensed data and the ground truthed data.

Once again if interest lies in estimating only the wheat acreage, the two-crop approach of Section 3 can be applied. Then an estimate of wheat proportion is obtained from (3.4) or (3.5) as the case may be, either first

obtaining it stratumwise and then combining as we did above in (5.4) and (5.5) or directly, depending upon whether or not the classification error matrix is the same for different strata. Subsequently, the precision of this estimate and the sample size necessary to achieve a desired precision with minimum cost can be easily obtained by applying our technique of Section 3.

Sample Size

Taking the cost factor into consideration, suppose we want to determine the sample size that either minimizes the sampling cost for a specified precision or maximizes the precision of the estimate for a fixed cost. Though a large initial cost is involved in acquiring remotely sensed data, presently we are mainly concerned with the cost of the processing and labeling of the sampled data. In general, any such cost can be considered as

$$C_t = c_1 m_t + c_2 m_t r + c_3 m_t rs$$

for the sample in stratum R_t, and

$$C = (c_1 + c_2 r + c_3 rs) \sum_{t=1}^{L} m_t$$

for the area of interest.

In case of unknown classification error matrix or matrices, there is an additional cost of sampling the ground truth, say C'. As such the total cost involved is $C = C + C'$. Now if the cost is fixed, say $C'' \leq C_0$, a determination of sample sizes for both the unlabeled remotely sensed data in all three categories and the ground truth for various crops can be achieved by solving equations obtained by equating the partial derivatives of

$$MSE(\hat{p}_k) + \lambda(C''-C_0) \ , \qquad k = 1,2,\ldots,m$$

where λ is a Lagrange multiplier, with respect to m_t, r, s and the ground truth sample sizes to zero. Similarly, when the MSE $(\hat{p}_k)$ is fixed, say σ_k^2, k=1,2,...,m, again this can be achieved by considering the function

$$C'' + \lambda_k[MSE(\hat{p}_k) - \sigma_k^2], \quad k = 1,2,\ldots,m$$

for minimization. This, of course, would lead to k different values for various sample sizes unless we consider the minimization from the point of a specific crop-type proportion estimate. On the other hand, a unique determination can be obtained by considering the largest value obtained in each case.

It may be pointed out that under this procedure, it will be difficult to give any closed form expression for any sample size and its carrying out would involve some optimization technique.

If the classification error matrix (or matrices) is known, the sample sizes m_t(t=1,2,...,L), r and s can be easily determined by minimizing $Var(\hat{e}(\pi_k)) + \lambda(C-C_0)$ or $C + \lambda_k[Var(\hat{e}(\pi_k)) - \sigma_k^2]$ as the case may be. Moreover, the sample size problem in the case of unknown classification error matrix or matrices can be treated either by assuming the classification errors known or by investigating the two types of sampling separately.

6. FURTHER REMARKS

In actual practice it may not be possible to have every data point identified with one of the crops in the area of interest, particularly if the area is large. This may be caused by not knowing all crop-types that exist in the area or some data points representing pixels falling on the field boundaries. As such the model developed in this report may be viewed somewhat restricted. Its use for performing a large area crop inventory may be considered subsequent to obtaining information about the agricultural practices in the area.

It is extremely difficult to model the problem of a large area crop inventory in its full generality unless certain constraints are imposed. The condition of identifiability is one such constraint that one must have in order to deal with the problem analytically.

APPENDIX 1

A.1. Variances of Components of $\hat{p}$

For $\hat{p}$ given in (2.5), the covariance matrix,

$$E[(\hat{p} - p)(\hat{p} - p)^T] = P^{-1} E[(\hat{e} - e)(\hat{e} - e)^T](P^{-1})^T$$

or

$$\sum_{\hat{p}} = (P^{-1}) V (P^{-1})^T \tag{A.1}$$

where V denotes the covariance matrix of $\hat{e}$. Denoting the $(i, j)^{th}$ element of P^{-1} by p^{ij}, it follows that the variance of $\hat{p}_i$, the ith element of $\hat{p}$, is given by

$$\text{Var } (\hat{p}_i) = \sum_{j=1}^{m} (p^{ij})^2 \text{ Var } (\hat{e}_j) + \sum_{\substack{j=1 \\ j \neq k}}^{m} \sum_{k=1}^{m} p^{ij} p^{ik} \text{ Cov } (\hat{e}_j, \hat{e}_k) \tag{A.2}$$

where $V(\hat{e}_j)$ and Cov $(\hat{e}_j, \hat{e}_k)$ would depend upon the sampling scheme used for obtaining samples of unlabeled remotely sensed data points.

In the case of random sampling with sampling unit as pixel (i.e. one data point),

$$\text{Var } (\hat{e}_j) = \frac{e_j(1 - e_j)}{n} \tag{A.3}$$

and

$$\text{Cov } (\hat{e}_j, \hat{e}_k) = - \frac{e_j e_k}{n}, \; j \neq k, \; j, k = 1, 2, .., m,$$

ignoring the finite population correction due to large population size. Next

an unbiased estimate of these quantities is given by

$$\widehat{\text{Var}}\ (\hat{e}_j) = \frac{\hat{e}_j(1 - \hat{e}_j)}{n - 1}$$

$$\widehat{\text{Cov}}\ (\hat{e}_j, \hat{e}_k) = -\frac{\hat{e}_j\,\hat{e}_k}{n - 1},\ j \neq k,\ j, k = 1, 2, .., m.$$

On the other hand if the sampling unit is a 5 x 6 mile segment consisting of r pixels then considering a random sample of m segments (here for the sample size one may consider n = mr data points) from the total of M segments in the area of interest and again ignoring the finite population correction, one gets (Cochran, 1963)

$$\text{Var}\ (\hat{e}_j) = \frac{1}{m(M - 1)} \sum_{i=1}^{M} (e_{ji} - e_j)^2$$

and

$$\text{Cov}\ (\hat{e}_j, \hat{e}_k) = \frac{1}{m\ (M-1)} \sum_{i=1}^{M} (e_{ji} - e_j)(e_{ki} - e_k),\ j \neq k \qquad (A.4)$$

$$j, k = 1, 2, .., m$$

where e_{ji} denotes the proportion of classified data points in π_j for the ith segment. Once again, for their unbiased estimates

$$\widehat{\text{Var}}\ (\hat{e}_j) = \frac{1}{m(m - 1)} \sum_{i=1}^{m} (\hat{e}_{ji} - \hat{e}_j)^2$$

and

$$\widehat{\text{Cov}}\ (\hat{e}_j, \hat{e}_k) = \frac{1}{m(m-1)} \sum_{i=1}^{m} (\hat{e}_{ji} - \hat{e}_j)(\hat{e}_{ki} - \hat{e}_k),\ i \neq k,$$

$$j, k = 1, 2, \ldots, m.$$

Similarly, components of V and their estimates can be obtained for other types of sampling plans. Making appropriate substitution in (A.1) or (A.2), variances for the components of $\hat{p}$ and their estimates are then obtained.

A.2. Mean Square Errors of Components of $\hat{\hat{p}}$.

First we calculate the bias of $\hat{\hat{p}}$ given by

$$\text{Bias}\ (\hat{\hat{p}}) = E[\hat{\hat{p}}-p]$$

$$= E\ [\hat{P}^{-1}\hat{e} - P^{-1}e]$$

$$= E\ [\hat{P}^{-1}(\hat{e}-e) + (\hat{P}^{-1} - P^{-1})e]$$

$$= E\ [\hat{P}^{-1} - P^{-1}]e \qquad (A.5)$$

because the first term is zero due $E(\hat{e}-e) = 0$ for a given $\hat{P}^{-1}$. Clearly, the bias depends upon how much bias there is in $\hat{P}^{-1}$, and

$$\text{Bias}\ (\hat{p}) = (\text{Bias}\ (\hat{P}^{-1}))e.$$

In order to find the mean square error of any component of $\hat{\hat{p}}$, let us first consider the evaluation of matrix,

$$E\ [(\hat{\hat{p}}-p)(\hat{\hat{p}}-p)^T] = E\ [(\hat{P}^{-1}\hat{e} - P^{-1}e)(\hat{P}^{-1}\hat{e} - P^{-1}e)^T]$$

so

$$= E\,[(\hat{P}^{-1})(\hat{e}-e)(\hat{e}-e)^T(\hat{P}^{-1})^T + (\hat{P}^{-1} - P^{-1})\; ee^T\; (\hat{P}^{-1} - P^{-1})^T]$$

$$= E\,[(\hat{P}^{-1})\; V\; (\hat{P}^{-1})^T] + E\,[(\hat{P}^{-1} - P^{-1})\; ee^T\; (\hat{P}^{-1} - P^{-1})^T] \qquad (A.6)$$

where E stands for expectation with respect to $\hat{P}$. Again, denoting the (i, j)th element of P^{-1} by P^{ij} and that of $\hat{P}^{-1}$ by $\hat{P}^{ij}$, it follows from (A.6) that the mean square error of $\hat{p}_i$, the i^{th} component of $\hat{p}$, is given by

$$MSE\,(\hat{\hat{p}}_i) = E\Bigg[\sum_{j=1}^{m} (\hat{P}^{ij})^2\; Var\,(\hat{e}_j) + \sum_{\substack{j=1 \\ j\neq k}}^{m}\sum_{k=1}^{m} \hat{P}^{ij}\,\hat{P}^{ik}\; Cov\,(\hat{e}_j, \hat{e}_k)$$

$$+ \sum_{j=1}^{m} e_j^2\; E\,[\hat{P}^{ij} - P^{ij}]^2 + \sum_{\substack{j=1 \\ j\neq k}}^{m}\sum_{k=1}^{m} e_j e_k E[(\hat{P}^{ij} - P^{ij})(\hat{P}^{ik} - P^{ik})]\Bigg], \qquad (A.7)$$

$i = 1, 2, \ldots, m.$

Once again, V, i.e. Var $(\hat{e}_j)$ and Cov $(\hat{e}_j, \hat{e}_k)$, j and k = 1, 2, ..., m, may be obtained as in (A.3) and (A.4). If some other sampling plan is used for selecting remotely sensed data to obtain the estimates $\hat{e}_j$'s, expression for V can accordingly be obtained. To evaluate expectation in (A.7), one needs to find the distribution of $\hat{P}$. This will, of course, depend upon how $\hat{P}$ is obtained. In general, it will be difficult to obtain any exact distribution of $\hat{P}$. However, if the sampling of ground truth involves separate independent samples from each crop and $\hat{P}$ is obtained as the matrix of observed proportions among randomly selected pixels classified

into different crops using a classifier, each column vector of $\hat{P}$ has a multinomial distribution and is stochaotically independent of the others in $\hat{P}$. Since expectation in (A.7) is for elements of $\hat{P}^{-1}$, it may not be easy to derive the MSE $(\hat{p}_i)$ in a closed form, especially if the number of crops is large.

APPENDIX 2

Two-Crop Case

First we derive the $MSE(\hat{p}_1)$ as in (3.8).

Proof of (3.8)

Considering the estimates $\hat{\Phi}_1$, $\hat{\Phi}_2$ and $\hat{e}$, being obtained from independent sets of samples, it follows by an application of the δ-method that

$$MSE(\hat{p}_1) \doteq \left(\frac{\partial p_1}{\partial e_1}\right)^2 Var(\hat{e}_1) + \left(\frac{\partial p_1}{\partial \Phi_1}\right)^2 Var(\hat{\Phi}_1) + \left(\frac{\partial p_1}{\partial \Phi_2}\right)^2 Var(\hat{\Phi}_2)$$

$$\doteq \frac{1}{(1-\Phi_1-\Phi_2)^2} Var(\hat{e}_1) + \left[\frac{e_1-1+\Phi_2}{(1-\Phi_1-\Phi_2)^2}\right]^2 Var(\hat{\Phi}_1)$$

$$+\left[\frac{e_1-\Phi_1}{(1-\Phi_1-\Phi_2)^2}\right]^2 Var(\hat{\Phi}_2) \quad .$$

Hence

$$MSE(\hat{p}_1) \doteq \frac{1}{(1-\Phi_1-\Phi_2)^2}\left(Var(\hat{e}_1) + \left[1-\frac{e_1-\Phi_1}{1-\Phi_1-\Phi_2}\right]^2 Var(\hat{\Phi}_1)\right.$$

$$\left.+\left[\frac{e_1-\Phi_1}{1-\Phi_1-\Phi_2}\right]^2 Var(\hat{\Phi})\right) .$$

Here dot with equality sign means equality with approximation. This establishes (3.8).

For a determination of sample size necessary to minimize the cost subject to $MSE(\hat{p}_1) \leq \sigma^2$ as discussed in section 3, it is achieved by minimizing the function

$$F = C(N) + \lambda\,(\mathrm{MSE}\,(\hat{\hat{p}}_1) - \sigma^2)$$

with respect to n, N_1 and N_2, where C (N) is given in (3.12) and $\mathrm{MSE}(\hat{\hat{p}}_1)$ is given in (3.10). By rewriting, we have

$$F = c_1 n + (c_1 + c_2)(N_1 + N_2) + \lambda\,(1 - \Phi_1 - \Phi_2)^{-2} \cdot \left[\frac{e_1(1-e_1)}{n} + (1-p_1)^2\,\frac{\Phi_1(1-\Phi_1)}{N_1} + p_1^2\,\frac{\Phi_2(1-\Phi_2)}{N_2} - \sigma^2(1-\Phi_1-\Phi_2)^2\right].$$

Taking partial derivatives of F with respect to n, N_1 and N_2 and equating each to zero, one obtains the following set of equations.

$$c_1 - \lambda(1-\Phi_1-\Phi_2)^{-2}\,\frac{e_1(1-e)}{n^2} = 0$$

$$(c_1 + c_2) - \lambda(1-\Phi_1-\Phi_2)^{-2}\,(1-p_1)^2\,\frac{\Phi_1(1-\Phi_1)}{N_1^2} = 0$$

$$(c_1 + c_2) - \lambda(1-\Phi_1-\Phi_2)^{-2}\,p_1^2\,\frac{\Phi_2(1-\Phi_2)}{N_2^2} = 0 \quad .$$

Considering only the admissible solution of these equations, one has

$$n = \sqrt{\frac{\lambda e_1(1-e_1)}{c_1(1-\Phi_1-\Phi_2)^2}}$$

$$N_1 = (1-p_1)\sqrt{\frac{\lambda\,\Phi_1(1-\Phi_1)}{(c_1+c_2)(1-\Phi_1-\Phi_2)^2}} \qquad \text{(A.8)}$$

$$N_2 = p_1\sqrt{\frac{\lambda\,\Phi_2(1-\Phi_2)}{(c_1+c_2)(1-\Phi_1-\Phi_2)^2}} \quad .$$

Considering that MSE $(\hat{\hat{p}}_1) = \sigma^2$ and making substitution in (3.10) for n, N_1 and N_2 obtained in (A.8), one gets

$$\sqrt{\lambda} = \frac{1}{\sigma^2(1-\Phi_1-\Phi_2)}\left[\sqrt{c_1 e_1(1-e_1)} + (1-p_1)\sqrt{(c_1+c_2)\,\Phi_1(1-\Phi_1)} + p_1\sqrt{(c_1+c_2)\Phi_2(1-\Phi_2)}\right] .$$

After substituting for $\sqrt{\lambda}$ in (A.8), the sample sizes n, N_1 and N_2 are obtained as following:

$$n = \sqrt{\frac{e_1(1-e_1)/c_1}{\sigma^2(1-\Phi_1-\Phi_2)^2}}\left[\sqrt{c_1 e_1\,(1-e_1)} + (1-p_1)\sqrt{(c_1+c_2)\Phi_1(1-\Phi_1)} + p_1\sqrt{(c_1+c_2)\Phi_2(1-\Phi_2)}\right]$$

$$N_1 = \frac{(1-p_1)}{\sigma^2(1-\Phi_1-\Phi_2)^2}\sqrt{\frac{\Phi_1(1-\Phi_1)}{c_1+c_2}}\left[\sqrt{c_1 e_1(1-e_1)} + (1-p_1)\sqrt{(c_1+c_2)\,\Phi_1(1-\Phi_1)} + p_1\sqrt{(c_1+c_2)\Phi_2(1-\Phi_2)}\right]$$

$$N_2 = \frac{p_1}{\sigma^2(1-\Phi_1-\Phi_2)^2}\sqrt{\frac{\Phi_2(1-\Phi_2)}{c_1+c_2}}\left[\sqrt{c_1 e_1\,(1-e_1)} + (1-p_1)\sqrt{(c_1+c_2)\Phi_1(1-\Phi_1)} + p_1\sqrt{(c_1+c_2)\Phi_2(1-\Phi_2)}\right] . \quad (A.9)$$

It can be easily seen that n is a monotone increasing and N_1, N_2 are monotone decreasing functions in c_2/c_1, the ratio of two types of cost. For when e_1, Φ_1, Φ_2, are unknown, estimates of n, N_1 and N_2 can be obtained from (A.9) by replacing these unknown quantities by their estimates.

APPENDIX 3

ERTS-1 DATA INVESTIGATION

FOR

WHEAT IDENTIFICATION

1. Hill County, Montana

   - Complete ground for evaluation in 2 × 6 mile area in Hill County North
   - Ground identifications of wheat, barley, oats in Hill County South
   - ERTS-1 data evaluated at three acquisition periods covering spring and winter wheat seasons

Date	Winter Wheat Stage	Spring Wheat Stage
May	Greening	Pre-emergence
June	Heading	100% cover
July	Mature	Headed

   - Classification performance results:

W - Spring/Winter Wheat
NW - Oats/Barley/Pasture

Commission/Omission Percentages

	W	NW
W	70	30
NW	20	80

May 23(t_1)

	W	NW
W	90	10
NW	15	85

June 27(t_2)

	W	NW
W	80	20
NW	5	95

July 16(t_3)

	W	NW
W	90	10
NW	5	95

May, June (t_1,t_2)

	W	NW
W	90	10
NW	5	95

May, July (t_1,t_3)

	W	NW
W	95	5
NW	0	100

May, June, July (t_1,t_2,t_3)

2. Burke County, North Dakota

- Complete ground truth for evaluation in 2 × 10 mile area
- ERTS-1 data evaluated at two acquisition periods

Date	Spring Wheat Stage
June 5	3"-4" growth
June 23	Jointing

- Classification performance results:

W - Spring Wheat
NW - Barley/Oats/Pasture/Summer Fallow

Commission/Omission Percentages

	W	NW
W	75	25
NW	10	90

June 5(t_1)

	W	NW
W	85	15
NW	10	90

June 23(t_2)

	W	NW
W	90	10
NW	5	95

June 5, June 23
(t_1,t_2)

REFERENCES

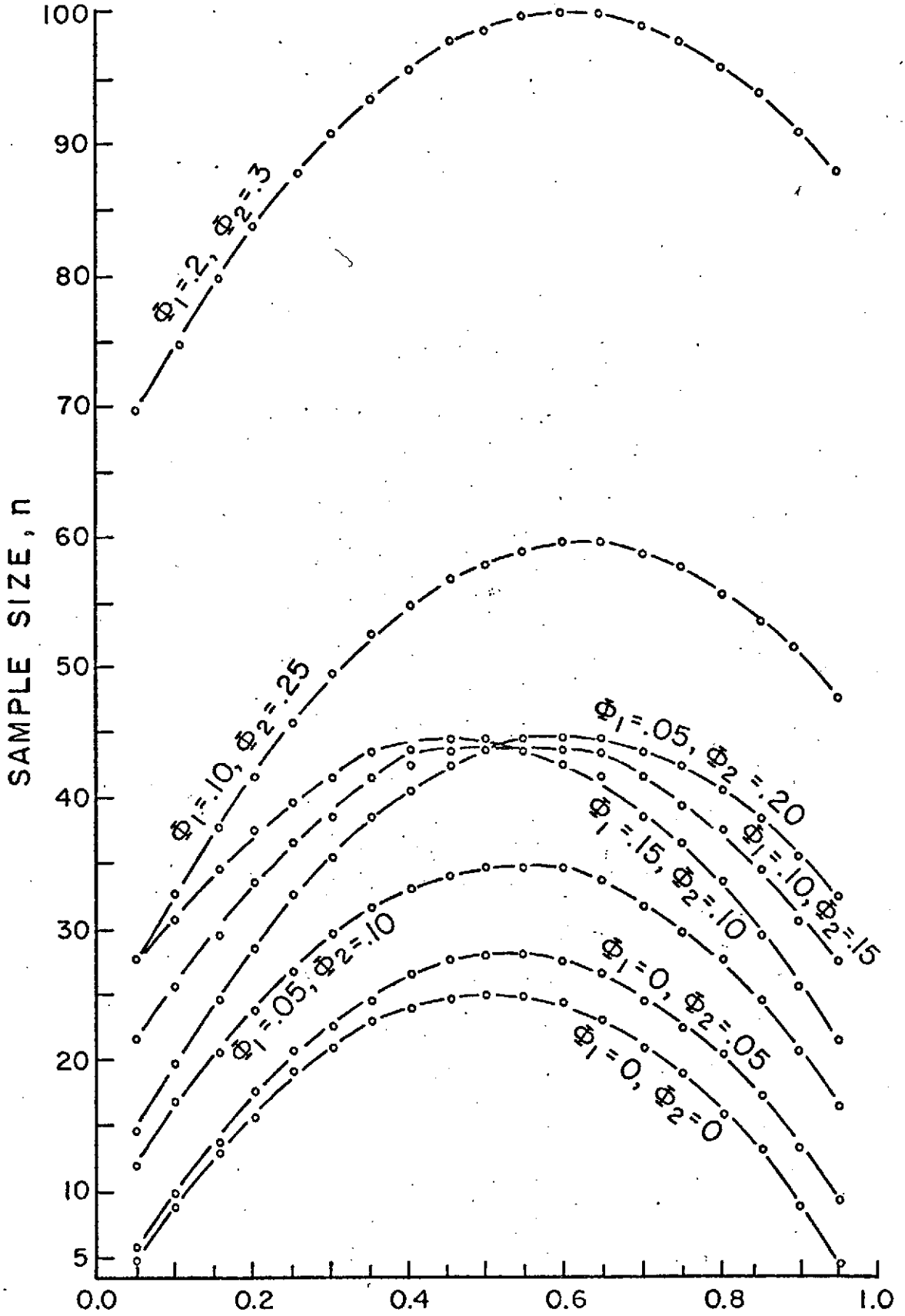
[1] Basu, J. P. and Odell, P. L. (1974): "On combining populations in statistical classification using remote sensing data," Technical Report for NASA.

[2] Chhikara, R. S. and Odell, P. L. (1973): "Acreage estimates for crops using remote sensing techniques," Technical Report for NASA.

[3] Chhikara, R. S. and Odell, P. L. (1974): "Estimation of crop acreage through sampling of remotely sensed data," Technical Report for NASA.

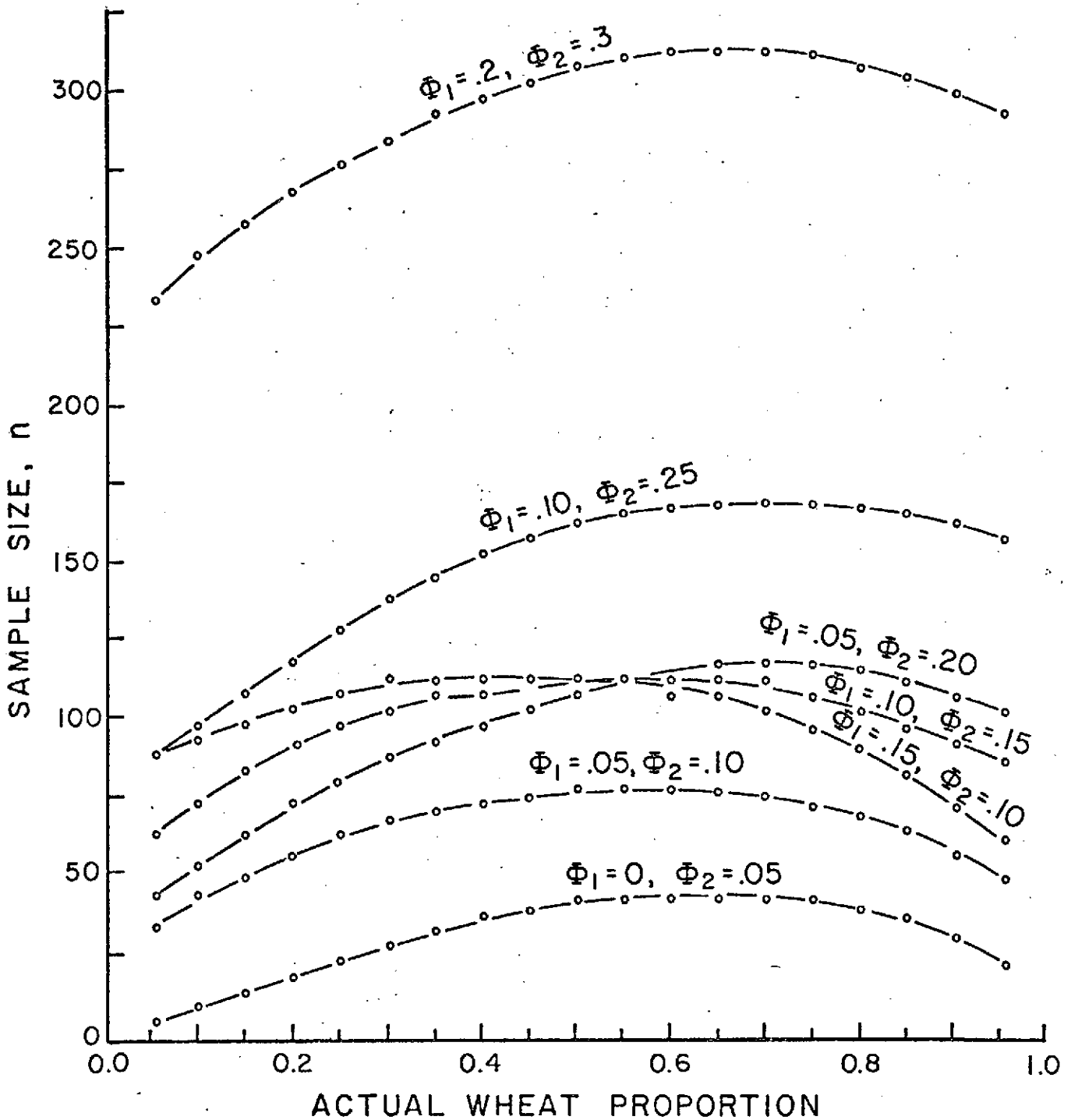
[4] Cochran, W. (1963): Sampling Techniques, 2nd Edition, John Wiley and Sons, Inc., New York.

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



Case: Φ_1, Φ_2 known

FIGURE 1



Case: Φ_1, Φ_2 unknown and $c_2/c_1 = 5$

FIGURE 2

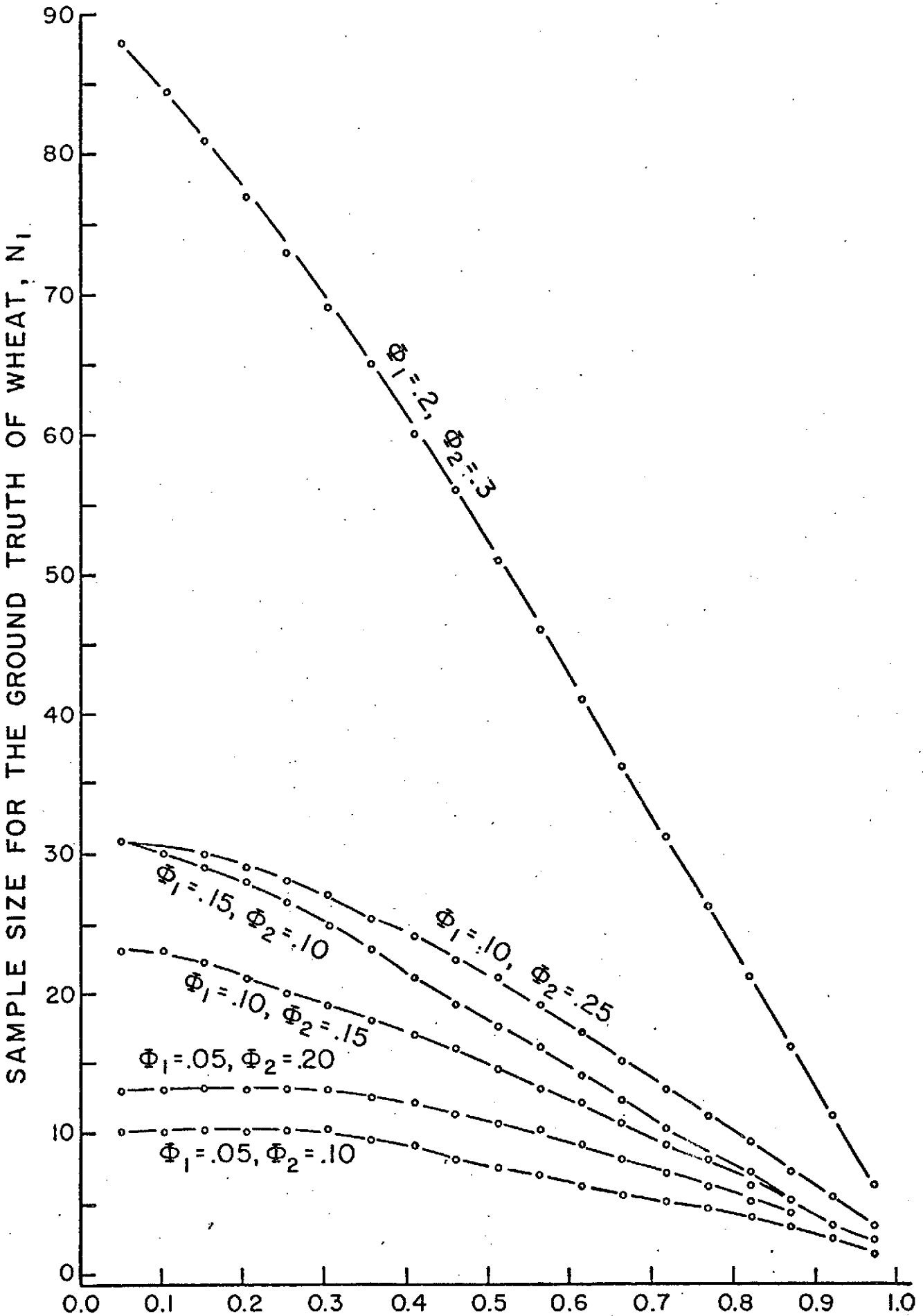


Figure 3

Case: Φ_1 , Φ_2 unknown and $c_2/c_1 = 5$

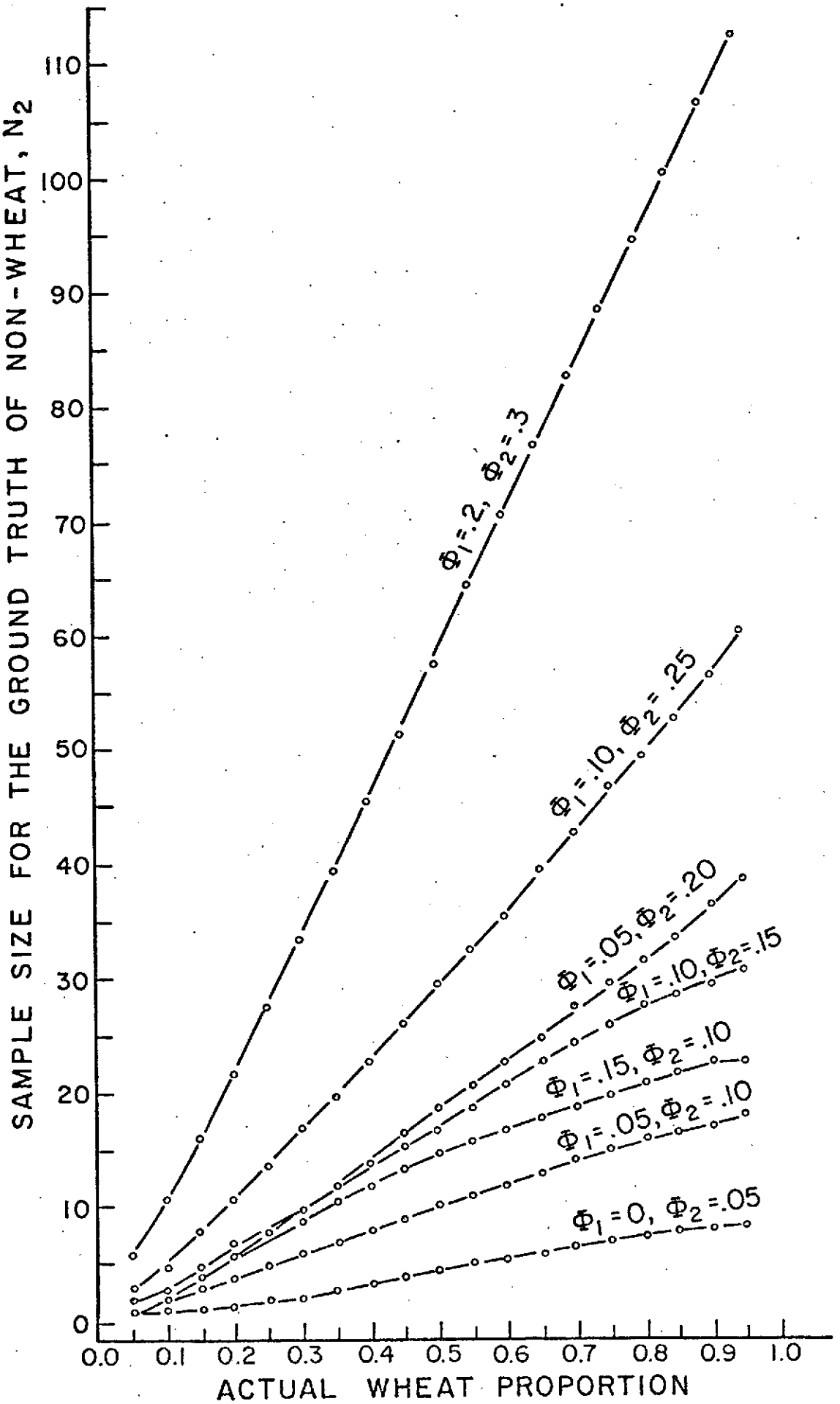
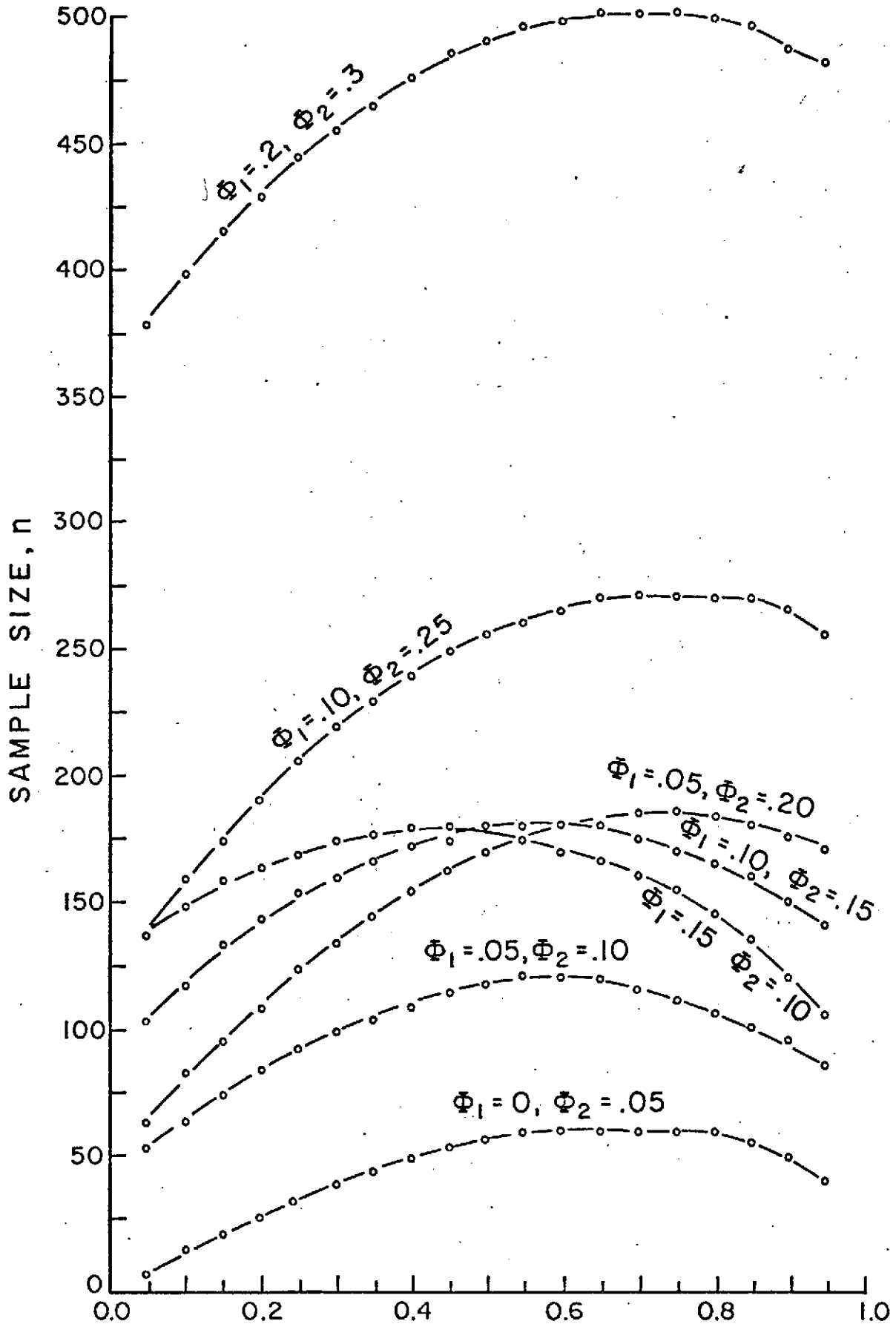


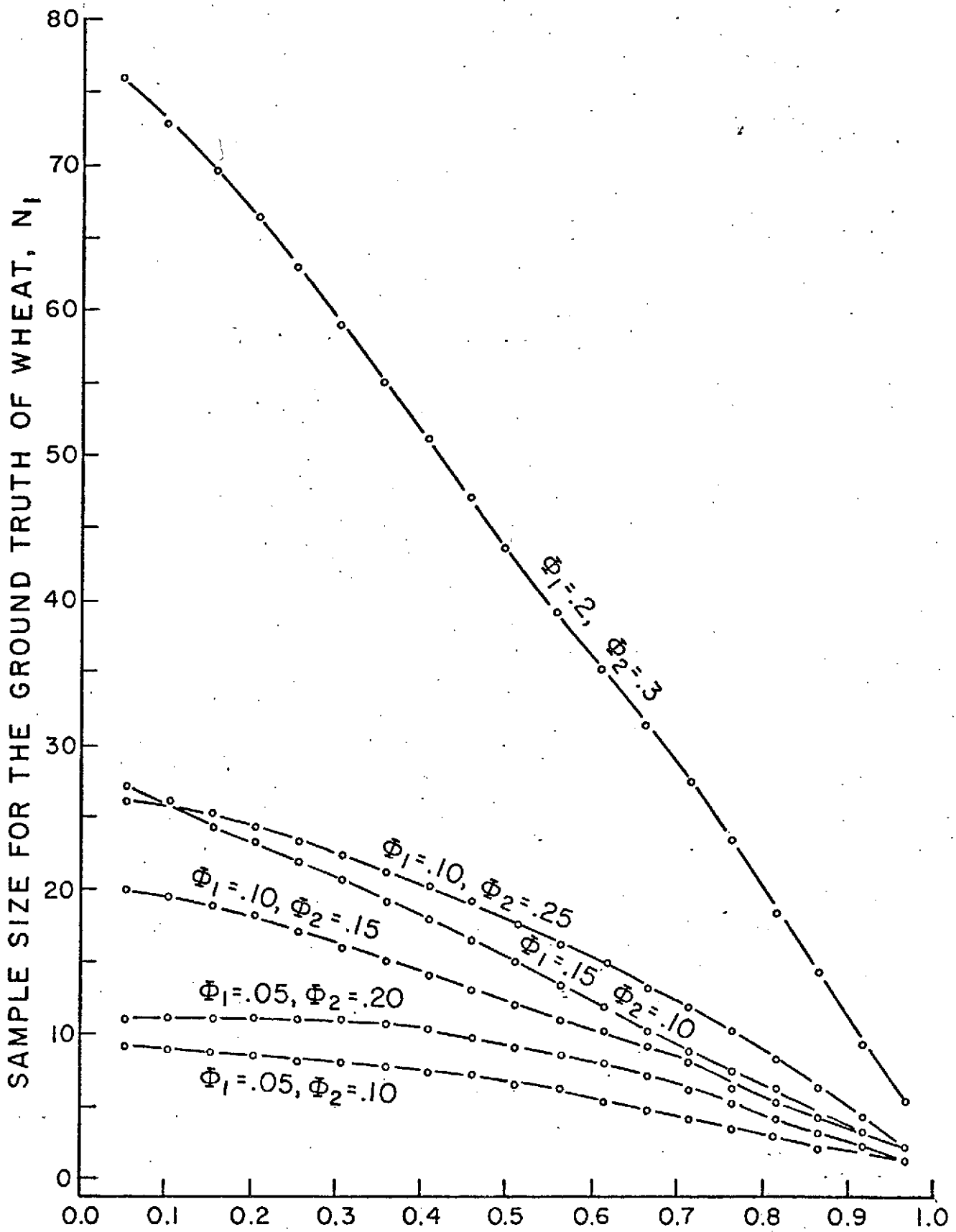
Figure 4

Case: Φ_1, Φ_2 unknown and $c_2/c_1 = 5$



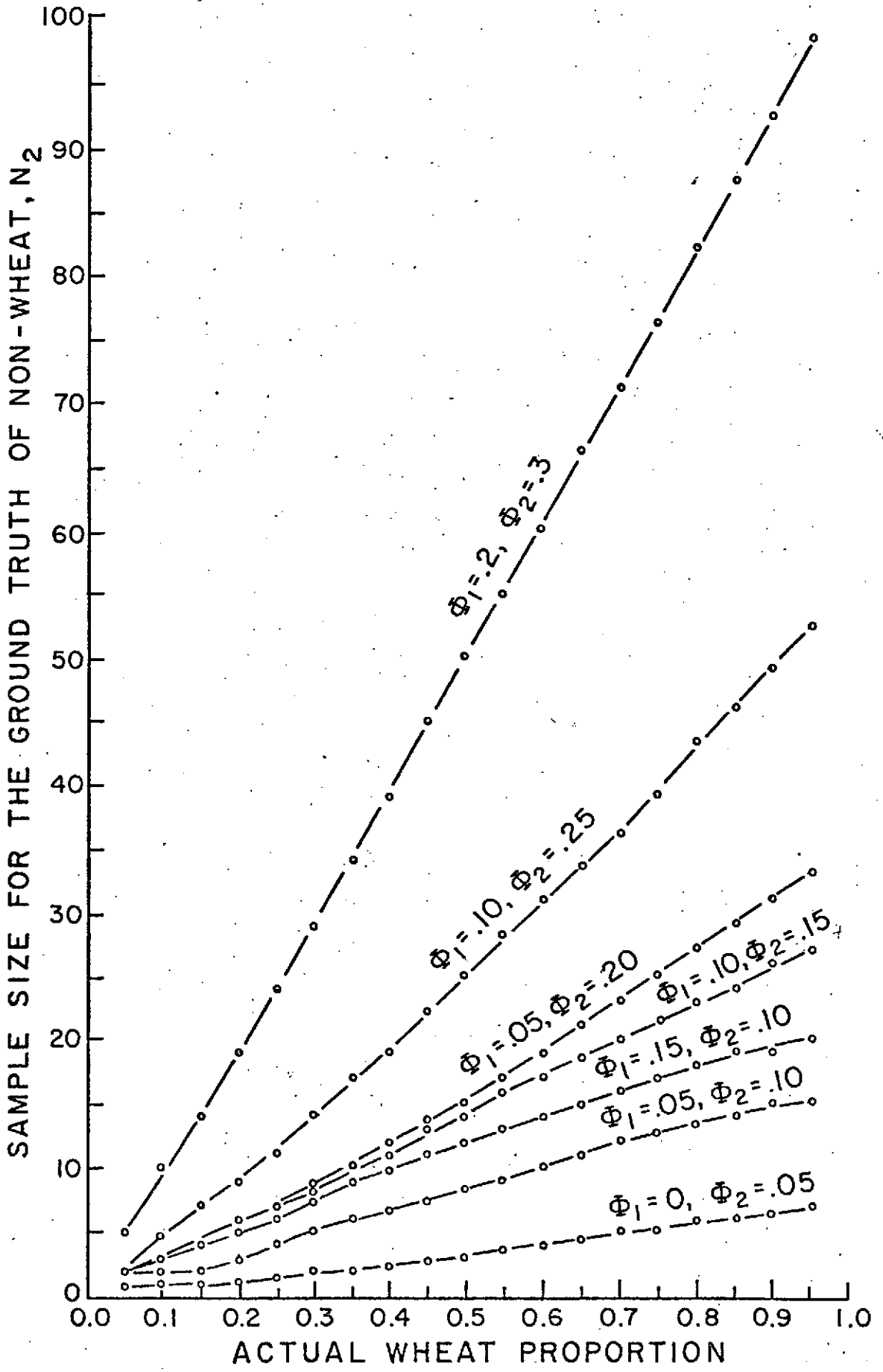
Case: Φ_1, Φ_2 unknown and $c_2/c_1 = 20$

Figure 5



Case: Φ_1, Φ_2 unknown and $c_2/c_1 = 20$

Figure 6



Case: Φ_1, Φ_2 unknown and $c_2/c_1 = 20$

Figure 7

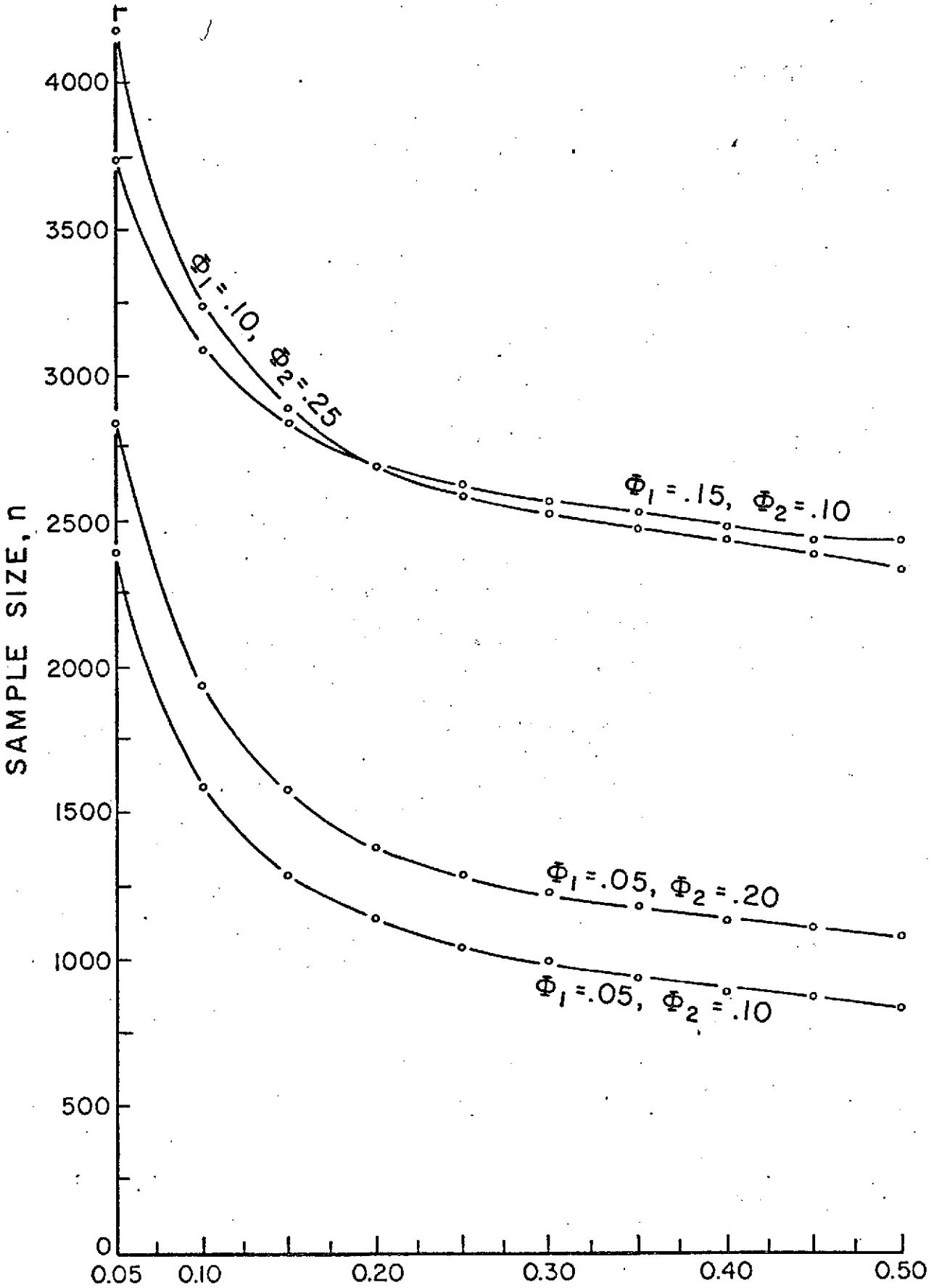
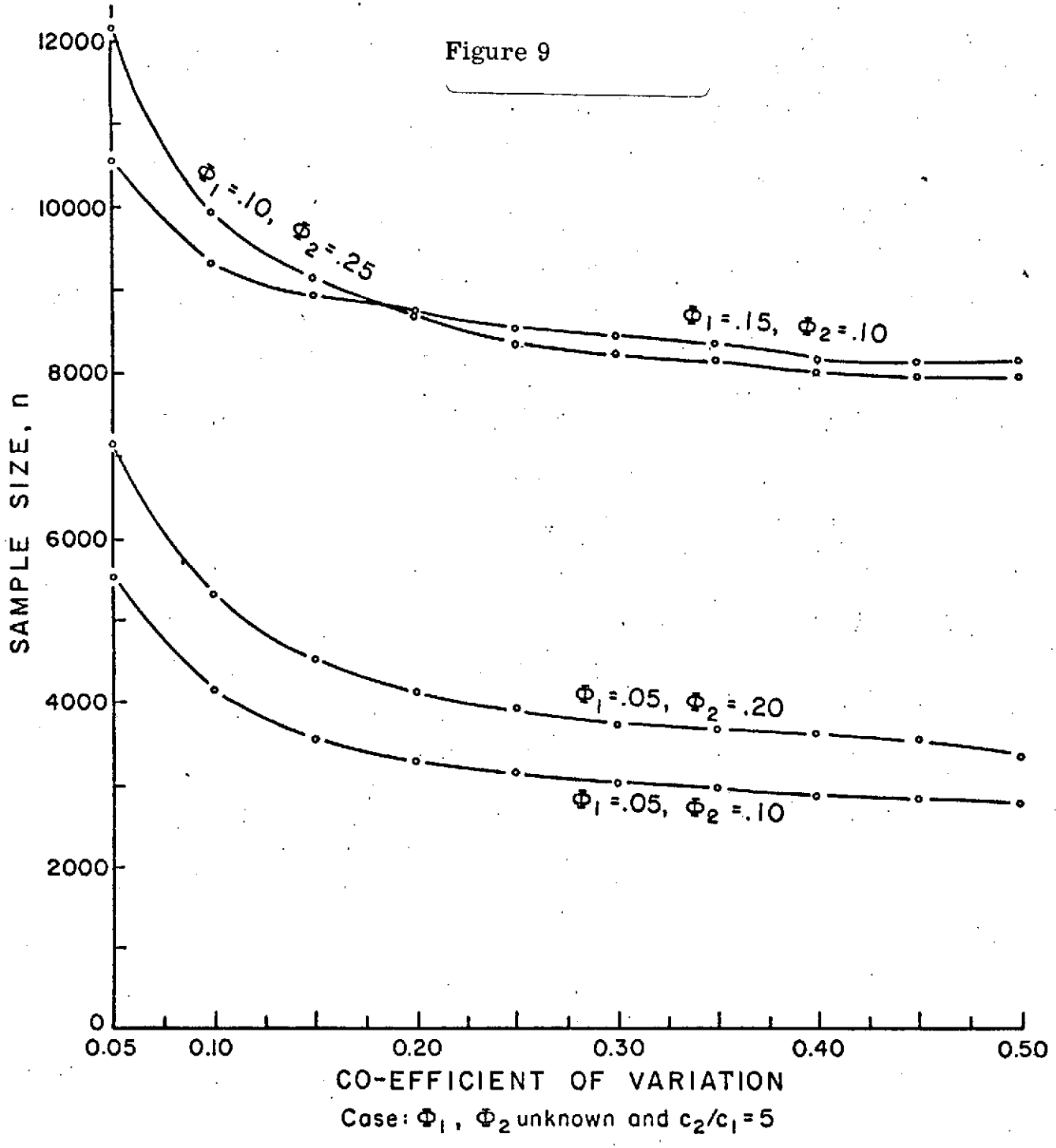


Figure 8

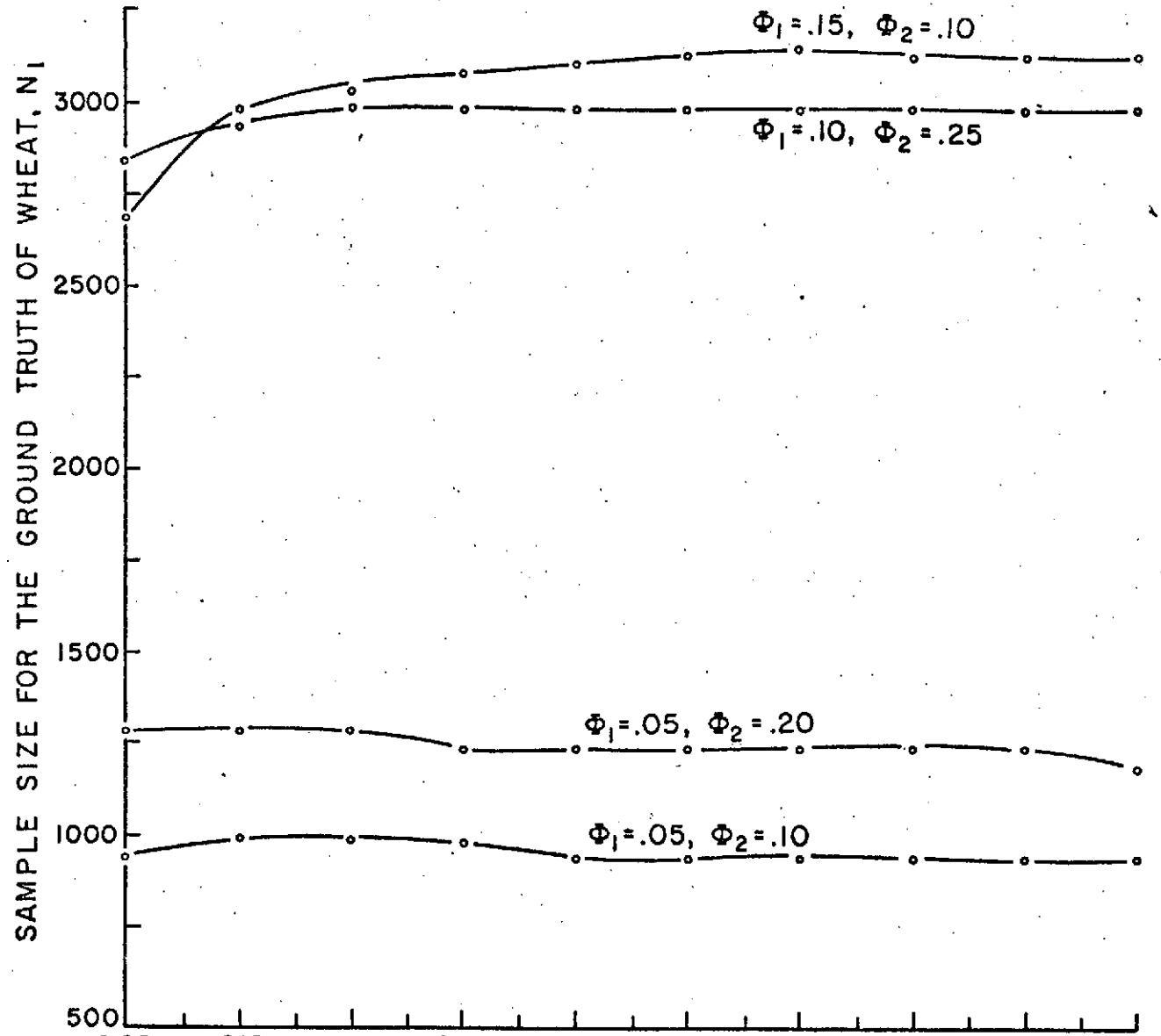
Case: Φ_1, Φ_2 known

Figure 9



Case: Φ_1, Φ_2 unknown and $c_2/c_1 = 5$

Figure 10



Case: Φ_1 , Φ_2 unknown and $c_2/c_1 = 5$

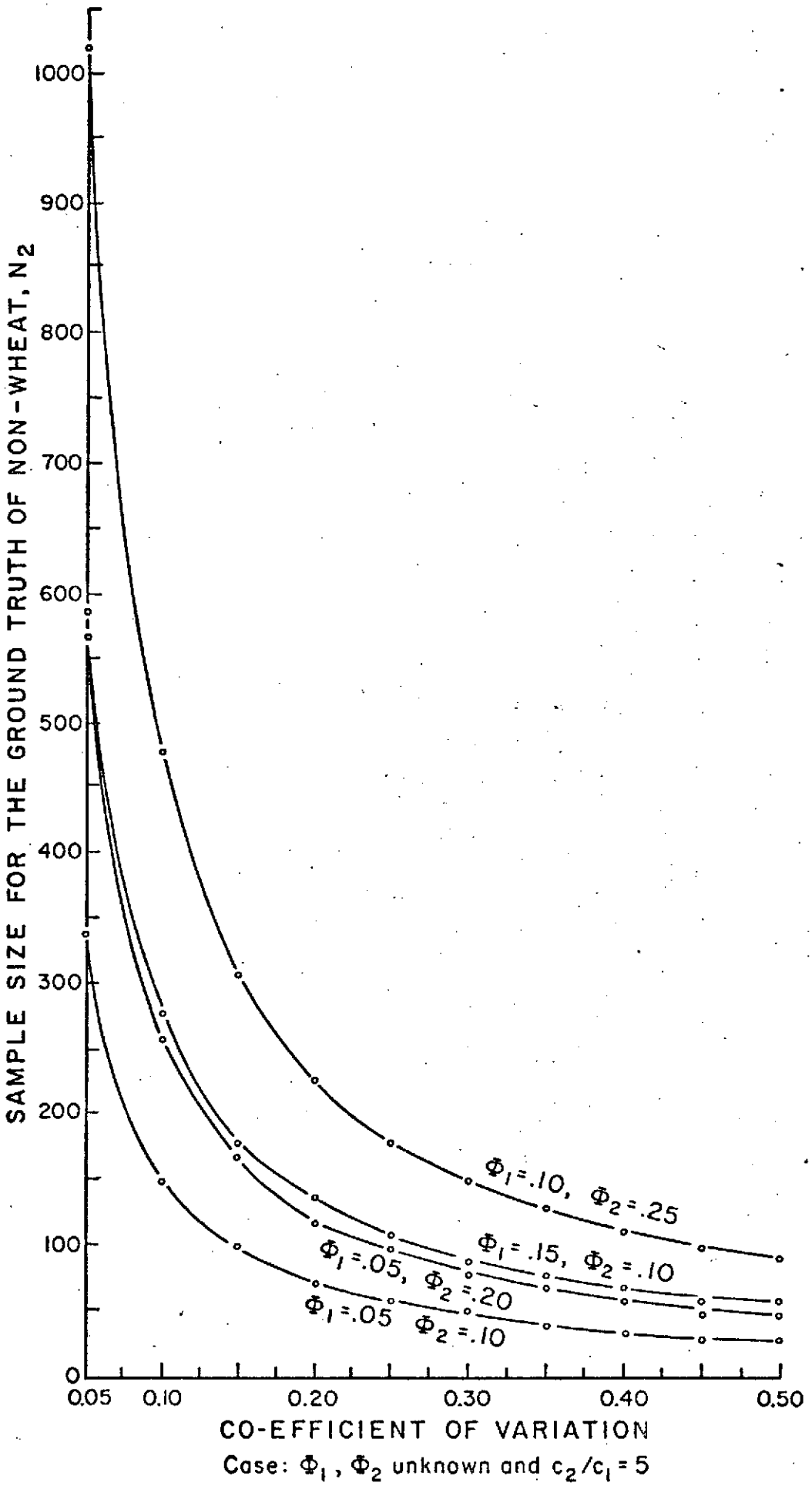
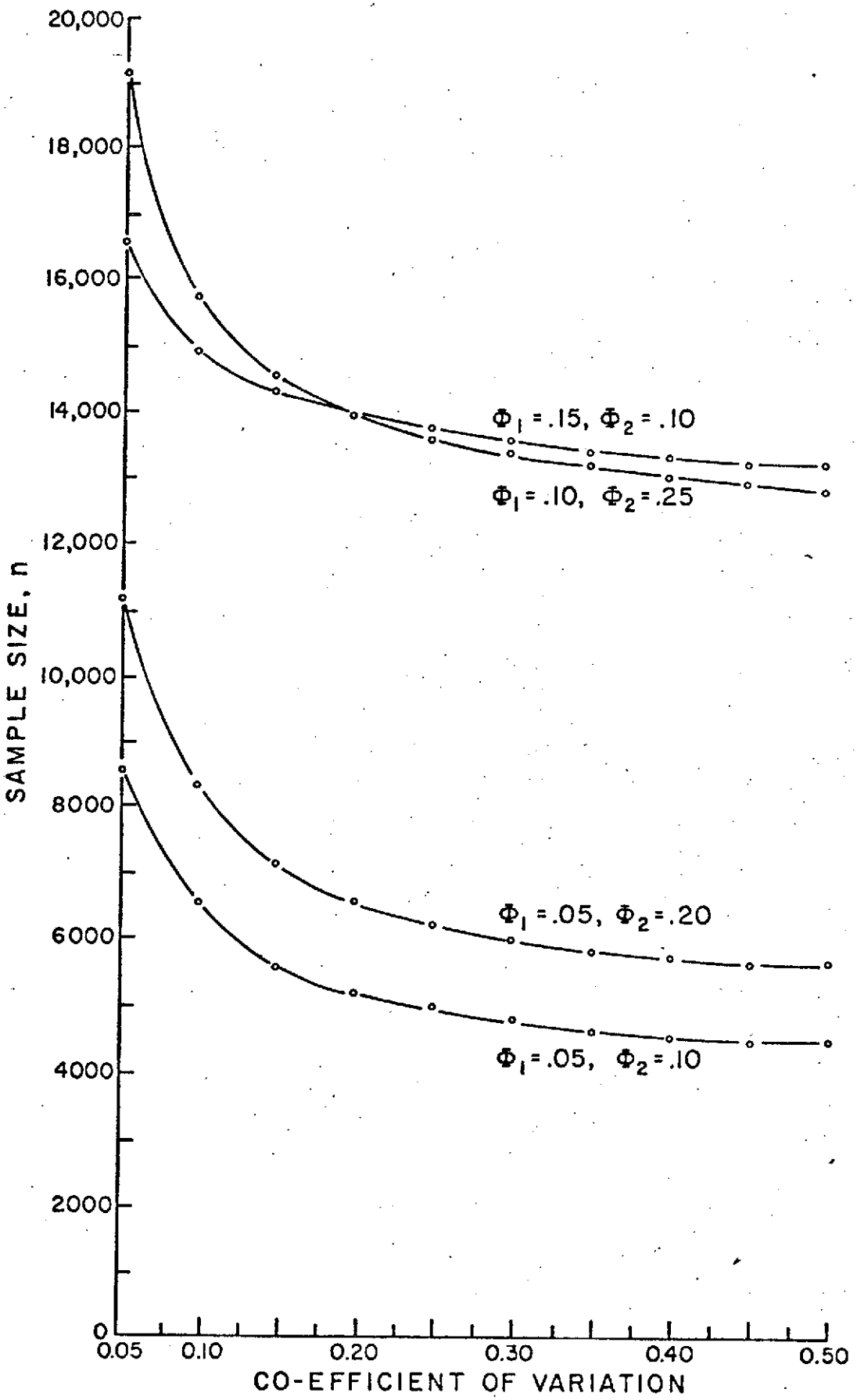

SAMPLE SIZE FOR THE GROUND TRUTH OF NON-WHEAT, N_2
1000
900
800
700
600
500
400
300
200
100
0
0.05
0.10
0.20
0.30
0.40
0.50
CO-EFFICIENT OF VARIATION
Φ_1 = .10, Φ_2 = .25
Φ_1 = .15, Φ_2 = .10
Φ_1 = .05, Φ_2 = .20
Φ_1 = .05 Φ_2 = .10
Case: Φ_1, Φ_2 unknown and $c_2/c_1 = 5$

FIGURE II



Case: Φ_1, Φ_2 unknown and $c_2/c_1 = 20$

FIGURE 12

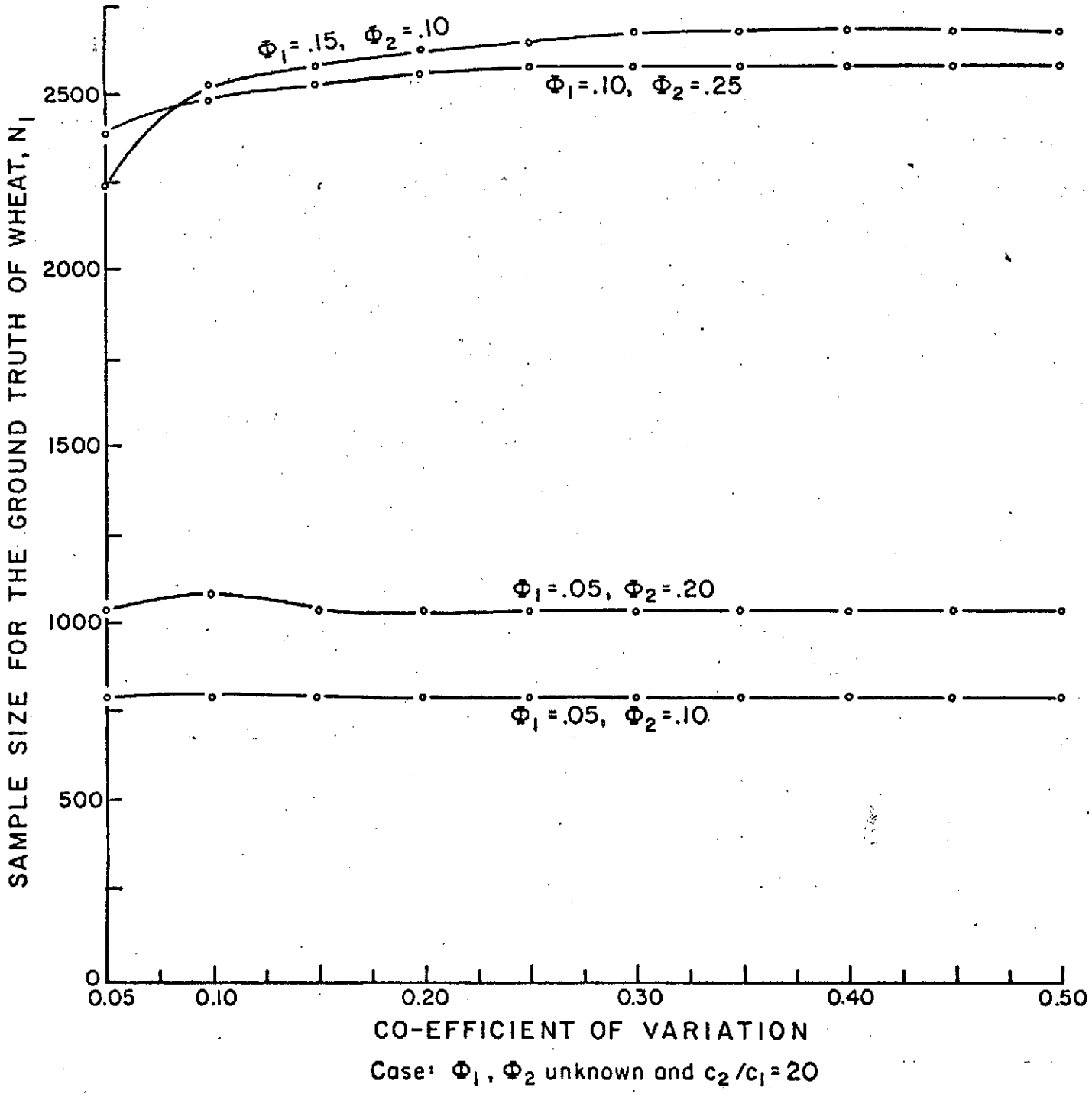
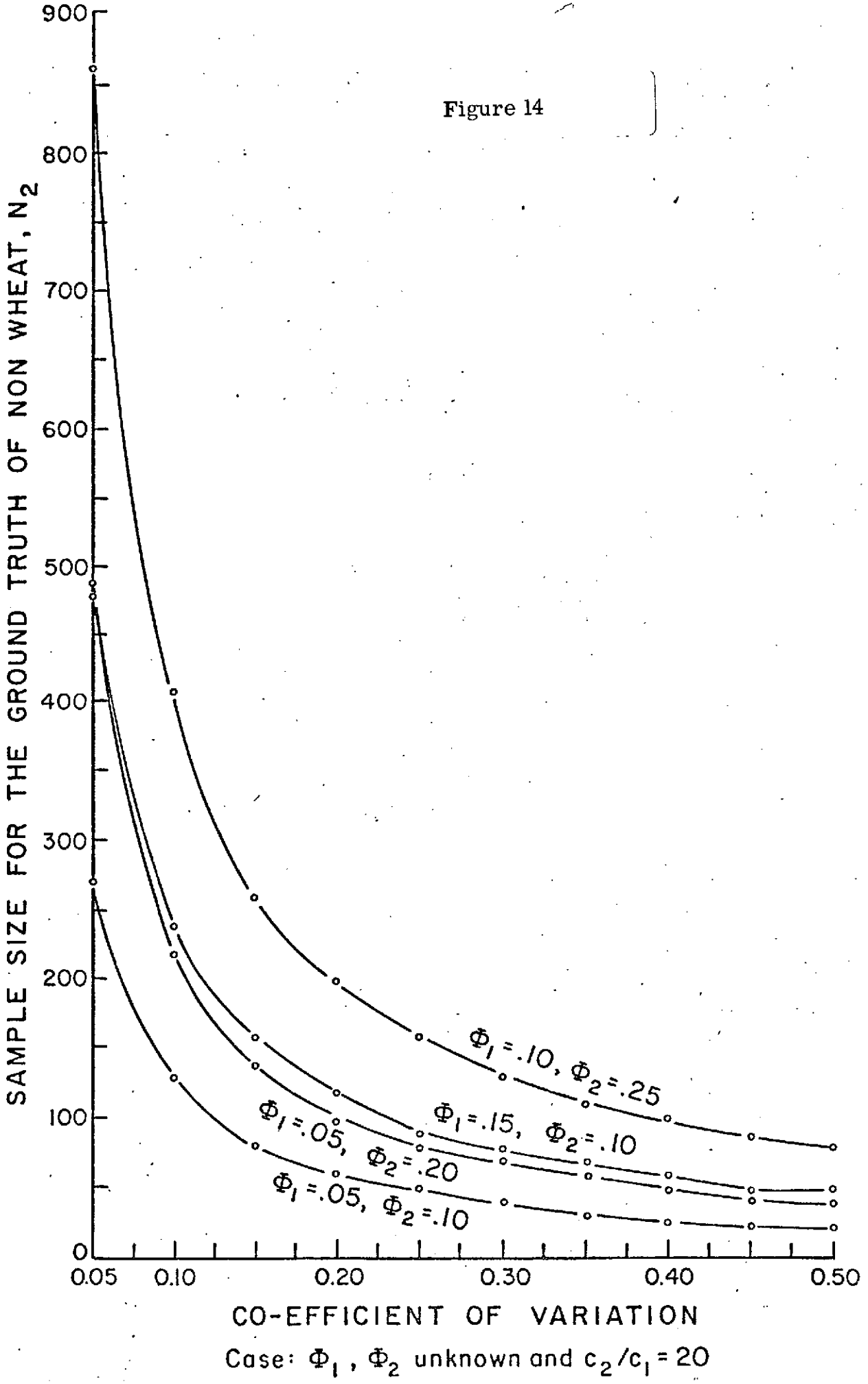

$\Phi_1 = .15,\ \Phi_2 = .10$
$\Phi_1 = .10,\ \Phi_2 = .25$
$\Phi_1 = .05,\ \Phi_2 = .20$
$\Phi_1 = .05,\ \Phi_2 = .10$
SAMPLE SIZE FOR THE GROUND TRUTH OF WHEAT, N_1
2500
2000
1500
1000
500
0
0.05
0.10
0.20
0.30
0.40
0.50
CO-EFFICIENT OF VARIATION
Case: Φ_1, Φ_2 unknown and $c_2/c_1 = 20$

Figure 13

Figure 14



Case: Φ_1 , Φ_2 unknown and $c_2/c_1 = 20$

TABLE : Sample sizes: n for the unlabeled remotely sensed data, and N_1 and N_2 for the ground truth of wheat and non-wheat, respectively, when precision for the proportion estimate is specified by $\sigma = .1$.

Actual wheat proportion P_1	Omission percentage Φ_1	Omission percentage Φ_2	Expected classified wheat proportion e_1	Sample size, n for Φ_1, Φ_2 known case	Sample sizes for Φ_1, Φ_2 unknown case: $c_2/c_1 = 5$ n	$c_2/c_1 = 5$ N_1	$c_2/c_1 = 5$ N_2	$c_2/c_1 = 10$ n	$c_2/c_1 = 10$ N_1	$c_2/c_1 = 10$ N_2
0.050	0.200	0.300	0.2250	70	235	88	6	379	76	5
0.050	0.100	0.250	0.1325	28	86	31	3	140	26	2
0.050	0.150	0.100	0.1875	28	88	31	2	140	27	2
0.050	0.100	0.150	0.1375	22	67	23	2	107	20	2
0.050	0.050	0.200	0.0875	15	43	13	2	67	11	2
0.050	0.050	0.100	0.0925	12	34	10	1	53	9	1
0.050	0.000	0.050	0.0475	6	6	0	1	7	0	1
0.100	0.200	0.300	0.2500	75	248	84	11	398	73	10
0.100	0.100	0.250	0.1650	33	101	30	5	159	26	5
0.100	0.150	0.100	0.2250	31	95	30	3	151	26	3
0.100	0.100	0.150	0.1750	26	77	23	3	121	19	3
0.100	0.050	0.200	0.1250	20	54	13	3	84	11	3
0.100	0.050	0.100	0.1350	17	43	10	2	66	9	2
0.100	0.000	0.050	0.0950	10	12	0	1	13	0	1
0.150	0.200	0.300	0.2750	80	259	81	17	415	69	14
0.150	0.100	0.250	0.1975	38	112	30	8	176	25	7
0.150	0.150	0.100	0.2625	35	102	29	5	160	24	4
0.150	0.100	0.150	0.2125	30	85	22	5	133	19	4
0.150	0.050	0.200	0.1625	25	64	14	5	98	11	4
0.150	0.050	0.100	0.1775	21	51	10	3	77	9	2
0.150	0.000	0.050	0.1425	14	17	0	1	20	0	1
0.200	0.200	0.300	0.3000	84	269	77	22	430	66	19
0.200	0.100	0.250	0.2300	42	122	29	11	191	24	9
0.200	0.150	0.100	0.3000	38	107	28	6	167	23	5
0.200	0.100	0.150	0.2500	34	93	21	7	144	18	6
0.200	0.050	0.200	0.2000	29	73	13	6	112	11	5
0.200	0.050	0.100	0.2200	24	59	10	4	86	8	3
0.200	0.000	0.050	0.1900	18	22	0	1	26	0	1
0.250	0.200	0.300	0.3250	88	278	73	28	444	63	24
0.250	0.100	0.250	0.2625	46	131	28	14	205	28	11
0.250	0.150	0.100	0.3375	40	111	25	8	172	22	6
0.250	0.100	0.150	0.2875	37	99	20	8	153	17	7
0.250	0.050	0.200	0.2375	33	82	13	8	124	11	7
0.250	0.050	0.100	0.2625	27	63	10	5	94	8	4
0.250	0.000	0.050	0.2375	21	27	0	2	32	0	1
0.300	0.200	0.300	0.3500	91	287	69	34	457	59	29
0.300	0.100	0.250	0.2950	50	140	27	17	218	22	14
0.300	0.150	0.100	0.3750	42	114	24	9	176	20	8
0.300	0.100	0.150	0.3250	39	104	19	10	161	16	8
0.300	0.050	0.200	0.2750	36	89	13	10	135	11	8
0.300	0.050	0.100	0.3050	30	68	10	6	101	8	5
0.300	0.000	0.050	0.2850	23	31	0	2	38	0	2
0.350	0.200	0.300	0.3750	94	294	65	40	467	55	34
0.350	0.100	0.250	0.3275	53	147	25	20	229	21	17
0.350	0.150	0.100	0.4125	44	116	23	11	179	19	9
0.350	0.100	0.150	0.3625	42	109	18	12	167	15	10
0.350	0.050	0.200	0.3125	39	90	12	12	145	10	10
0.350	0.050	0.100	0.3475	32	72	9	7	106	7	6
0.350	0.000	0.050	0.3325	25	35	0	3	43	0	2
0.400	0.200	0.300	0.4000	96	300	60	46	477	51	39
0.400	0.100	0.250	0.3600	55	153	24	23	239	20	19
0.400	0.150	0.100	0.4500	44	117	21	12	180	17	10
0.400	0.100	0.150	0.4000	42	112	17	14	172	14	11
0.400	0.050	0.200	0.3500	41	101	12	14	154	10	12
0.400	0.050	0.100	0.3900	33	75	9	8	111	7	6
0.400	0.000	0.050	0.3800	27	38	0	3	48	0	2
0.450	0.200	0.300	0.4250	98	305	56	52	454	47	45
0.450	0.100	0.250	0.3925	57	159	22	26	248	19	22
0.450	0.150	0.100	0.4875	45	117	19	13	180	16	11
0.450	0.100	0.150	0.4375	44	115	16	10	176	13	13
0.450	0.050	0.200	0.3875	43	106	11	15	162	9	13
0.450	0.050	0.100	0.4325	34	77	6	9	115	7	7
0.450	0.000	0.050	0.4275	28	41	0	4	52	0	3

p_1	ϕ_1	Φ_1	e_1	n	n	N_1	N_2	n	N_1	N_2
0.500	0.200	0.300	0.4500	99	309	51	58	491	44	50
0.500	0.100	0.250	0.4250	58	163	21	30	255	17	25
0.500	0.150	0.100	0.5250	45	116	17	15	178	14	12
0.500	0.100	0.150	0.4750	45	116	15	17	178	12	14
0.500	0.050	0.200	0.4250	44	110	10	19	168	9	15
0.500	0.050	0.100	0.4750	35	79	7	10	117	6	8
0.500	0.000	0.050	0.4750	28	43	0	4	56	0	3
0.550	0.200	0.300	0.4750	100	312	46	65	496	39	55
0.550	0.100	0.250	0.4575	59	167	19	33	261	16	28
0.550	0.150	0.100	0.5625	44	115	16	16	176	13	13
0.550	0.100	0.150	0.5125	45	117	13	19	180	11	16
0.550	0.050	0.200	0.4625	45	114	10	21	174	8	17
0.550	0.050	0.100	0.5175	35	80	7	11	118	6	9
0.550	0.000	0.050	0.5225	28	44	0	5	59	0	4
0.600	0.200	0.300	0.5000	100	314	41	71	499	35	60
0.600	0.100	0.250	0.4900	60	170	17	36	266	14	31
0.600	0.150	0.100	0.6000	43	112	14	17	172	11	14
0.600	0.100	0.150	0.5500	44	117	12	21	180	10	17
0.600	0.050	0.200	0.5000	45	116	9	23	178	7	19
0.600	0.050	0.100	0.5600	35	80	6	12	119	5	10
0.600	0.000	0.050	0.5700	28	45	0	5	61	0	4
0.650	0.200	0.300	0.5250	100	314	36	77	501	31	66
0.650	0.100	0.250	0.5225	60	171	15	40	269	13	34
0.650	0.150	0.100	0.6375	42	109	12	18	167	10	15
0.650	0.100	0.150	0.5875	44	116	11	23	179	9	19
0.650	0.050	0.200	0.5375	45	118	8	25	181	7	21
0.650	0.050	0.100	0.6025	34	79	5	13	118	4	11
0.650	0.000	0.050	0.6175	27	45	0	6	62	0	4
0.700	0.200	0.300	0.5500	99	314	31	83	501	27	71
0.700	0.100	0.250	0.5550	59	172	13	43	271	11	36
0.700	0.150	0.100	0.6750	39	104	10	19	161	8	16
0.700	0.100	0.150	0.6250	42	114	9	24	176	8	20
0.700	0.050	0.200	0.5750	44	118	7	28	183	6	23
0.700	0.050	0.100	0.6450	32	77	5	14	116	4	12
0.700	0.000	0.050	0.6650	25	45	0	6	62	0	5
0.750	0.200	0.300	0.5750	98	313	26	89	500	23	76
0.750	0.100	0.250	0.5875	58	172	11	47	271	10	39
0.750	0.150	0.100	0.7125	37	99	8	20	153	7	17
0.750	0.100	0.150	0.6625	40	111	8	26	172	6	22
0.750	0.050	0.200	0.6125	43	118	6	30	183	5	25
0.750	0.050	0.100	0.6875	30	74	4	15	112	3	12
0.750	0.000	0.050	0.7125	23	43	0	7	61	0	5
0.800	0.200	0.300	0.6000	96	311	21	95	498	18	82
0.800	0.100	0.250	0.6200	56	171	9	50	270	8	43
0.800	0.150	0.100	0.7500	34	93	7	21	144	6	18
0.800	0.100	0.150	0.7000	38	107	6	28	167	5	23
0.800	0.050	0.200	0.6500	41	116	5	32	182	4	27
0.800	0.050	0.100	0.7300	26	70	3	16	108	3	13
0.800	0.000	0.050	0.7600	21	41	0	7	59	0	6
0.850	0.200	0.300	0.6250	94	307	16	101	493	14	87
0.850	0.100	0.250	0.6525	54	168	7	53	268	6	46
0.850	0.150	0.100	0.7875	30	85	5	22	133	4	19
0.850	0.100	0.150	0.7375	35	102	5	29	160	4	24
0.850	0.050	0.200	0.6875	39	114	4	34	179	3	29
0.850	0.050	0.100	0.7725	25	66	3	17	101	2	14
0.850	0.000	0.050	0.8075	18	38	0	8	55	0	6
0.900	0.200	0.300	0.6500	91	303	11	107	487	9	92
0.900	0.100	0.250	0.6850	52	165	5	57	263	4	49
0.900	0.150	0.100	0.8250	26	77	3	23	121	3	19
0.900	0.100	0.150	0.7750	31	95	3	30	151	3	26
0.900	0.050	0.200	0.7250	30	110	3	37	175	2	31
0.900	0.050	0.100	0.8150	21	60	2	17	93	2	15
0.900	0.000	0.050	0.8550	14	33	0	8	49	0	6
0.950	0.200	0.300	0.6750	88	297	6	113	479	5	98
0.950	0.100	0.250	0.7175	48	160	3	60	257	2	52
0.950	0.150	0.100	0.8625	22	67	2	23	107	2	20
0.950	0.100	0.150	0.8125	28	88	2	31	140	2	27
0.950	0.050	0.200	0.7625	33	105	2	39	168	1	33
0.950	0.050	0.100	0.8575	17	52	1	18	83	1	15
0.950	0.000	0.050	0.9025	10	27	0	8	41	0	7

ESTIMATION OF OPTIMUM ERRORS OF CLASSIFICATION FOR UNIVARIATE NORMAL POPULATIONS

by

Raj S. Chhikara

and

Patrick L. Odell

1. Introduction

For the multivariate normal populations, the problem of classification using the Bayes' discriminant procedure involves certain difficulties as to an exact evaluation of actual errors of classification, i.e. optimum probabilities of misclassification, and an optimal estimation of these errors, particularly in the case of small samples. If the populations are assumed to be univariate normal, these errors of classification are easily expressible in a close form. Yet even for this simplified case the problem of their estimation has not been fully examined. For example, to the authors' best knowledge no minimum variance unbiased estimates of these errors are given so far in the literature.

Recently Sorum [8] investigated the estimation of both expected and optimum errors of classification associated with the linear discriminant function for univariate normal populations with known common variance. For the optimum errors of classification, she considered the maximum likelihood estimates and their various slight variants in her study involving large samples. Hill [6] too considered the same problem but his investigation was primarily motivated towards examining expected errors of classification.

For a brief statement of the problem, let π_1 and π_2 be two normally distributed populations with means μ_1 and μ_2, respectively, and common variance σ^2. Having obtained an observation x from one of the two populations, the problem is to identify the population to which it belongs. Assuming equal a priori probabilities for π_1, π_2 and a constant cost of misclassifying any observation, the Bayes' discriminant criterion amounts to: classify the observation x into π_1 if

$$x \geq \frac{\mu_1+\mu_2}{2} \quad \text{when } \mu_1 - \mu_2 > 0$$

and

$$x \leq \frac{\mu_1+\mu_2}{2} \quad \text{when } \mu_1 - \mu_2 < 0,$$

otherwise classify x into π_2. Without loss of generality, we assume $\mu_1 - \mu_2 > 0$.

Suppose $P(i|j)$ denotes the probability of classifying an observation x from π_j into π_i, i and j=1,2. Then the actual errors and non-errors of classification under the above discriminant rule are

$$P(i|j) = \begin{cases} \Phi(-\frac{\Delta}{2}) & \text{if } i \neq j \\ 1-\Phi(-\frac{\Delta}{2}), & \text{otherwise} \end{cases} \tag{1.1}$$

Sometimes it is desirable to have the knowledge of $P(i|j)$'s. It provides an assessment of the confusion likely to arise between individuals of two groups and thereby it is helpful in correcting for bias, etc. For example, see Cochran [3] who deals with several aspects of statistical inference in presence of certain types of classification errors.

In this paper we consider the problem of estimating $P(i|j)$, i and j=1,2 when sample observations are available from the populations. In view of (1.1), one, however, only needs to consider the estimation of $\Phi(-\Delta/2)$ when an estimation of $P(i|j)$'s is desired.

Let $x_1, x_2, \ldots, x_{n_1}$ be a random sample from population π_1 and $y_1, y_2, \ldots, y_{n_2}$ be another random sample from population π_2. Considering the two parametric cases (i) μ_1, μ_2 unknown and σ^2 known and (ii) μ_1, μ_2 and σ^2 all unknown, the maximum likelihood estimate (MLE) of $\Phi(-\Delta/2)$ is $\Phi(-\hat{\Delta}/2)$ where $\hat{\Delta} = (\bar{x}-\bar{y})/\sigma$ in case (i) and $\hat{\Delta} = (\bar{x}-\bar{y})/s$ in case (ii); here $\bar{x}$ and $\bar{y}$ denote the two sample means and s is obtained by

$$(n_1+n_2-2)s^2 = \sum_1^{n_1} (x_i-\bar{x})^2 + \sum_1^{n_2} (y_i-\bar{y})^2 \quad .$$

Below in section 2 we obtain the minimum variance unbiased estimate (MVUE) of $\Phi(-\Delta/2)$ and then compare it with the MLE in section 3 by evaluating relative efficiencies of both MVUE and MLE with respect to the Rao-Cramer lower bound for variances of unbiased estimates of $\Phi(-\Delta/2)$. It can be shown that the Rao-Cramer lower bound is

$$\frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \phi^2\left(-\frac{\Delta}{2}\right) \tag{1.2}$$

in case (i) and

$$\frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{\Delta^2}{2(n_1+n_2-1)}\right) \phi^2\left(-\frac{\Delta}{2}\right) \tag{1.3}$$

in case (ii), where ϕ denotes the standard normal density function. The relative efficiency of an estimate is obtained by dividing the appropriate lower bound by the mean square error (MSE) of the estimate.

2. MVUE of $\Phi(-\frac{\Delta}{2})$

(i) μ_1, μ_2 unknown and σ^2 known: The random variable

$$\frac{1}{2}\left(Z\sqrt{4-\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}+\frac{\bar{x}-\bar{y}}{\sigma}\right),$$

where Z is the standard normal variate, is distributed normally with mean $\frac{\Delta}{2}$ and variance one. As a result

$$\Phi(-\tfrac{\Delta}{2}) = \text{Prob}\left(Z\sqrt{4-\left(\tfrac{1}{n_1}+\tfrac{1}{n_2}\right)}+\frac{\bar{x}-\bar{y}}{\sigma} \leq 0\right)$$

$$= E\left[\text{Prob}\left(Z \leq -\frac{\bar{x}-\bar{y}}{\sigma\sqrt{4-\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \mid \bar{x},\bar{y}\right)\right]$$

$$= E\left[\Phi\left(-\frac{\bar{x}-\bar{y}}{\sigma\sqrt{4-\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right)\right],$$

E denoting expectation with respect to $\bar{x},\bar{y}$, a set of complete sufficient statistics. Thus by the Rao-Blackwell theorem we have the MVUE of $\Phi(-\frac{\Delta}{2})$,

$$\hat{\Phi}(-\tfrac{\Delta}{2}) = \Phi\left(-\frac{\bar{x}-\bar{y}}{\sigma\sqrt{4-\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right). \qquad (2.1)$$

(ii) μ_1, μ_2 and σ^2, all unknown: Let U be a beta variate, $\beta(\frac{\nu-1}{2},\frac{\nu-1}{2})$, stochastically independent of s^2 which is $\sigma^2\chi^2/\nu$ distributed with $\nu = (n_1+n_2-2)$ degrees of freedom (d.f.). Then by applying theorem 1 in Ellison [5], it follows that $(2U-1)\sqrt{\nu}\, s/\sigma$ has the standard normal distribution. Furthermore, the random variable

$$\frac{1}{2\sigma}\left[(2U-1)\, s\sqrt{\nu\left(4-\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)}+(\bar{x}-\bar{y})\right]$$

has the normal distribution with mean $\Delta/2$ and variance one. Accordingly,

$$\Phi(-\tfrac{\Delta}{2}) = \text{Prob}\ \{(2U-1)s\sqrt{\nu[4-(\tfrac{1}{n_1}+\tfrac{1}{n_2}]} + (\bar{x}-\bar{y}) \leq 0\}$$

$$= E[\text{Prob}\{(2U-1)s\sqrt{\nu[4-(\tfrac{1}{n_1}+\tfrac{1}{n_2})]} + (\bar{x}-\bar{y}) \leq 0 \,|\, \bar{x},\bar{y},s\}]$$

where E denotes expectation with respect to the set of complete sufficient statistics $\bar{x}$, $\bar{y}$ and s. Thus the MVUE of $\Phi(-\frac{\Delta}{2})$ is

$$\hat{\Phi}(-\tfrac{\Delta}{2}) = \text{Prob}\{(2U-1)\ s\sqrt{\nu[4-(\tfrac{1}{n_1}+\tfrac{1}{n_2})]} + (\bar{x}-\bar{y}) \leq 0 \,|\, \bar{x},\ \bar{y},\ s\}\ .$$

$$= \text{Prob}\ \{U \leq \tfrac{1}{2} - (\bar{x}-\bar{y})/2s\sqrt{\nu[4-(\tfrac{1}{n_1}+\tfrac{1}{n_2})]}\ \,|\bar{x},\bar{y},s\} \qquad (2.2)$$

where U has a $\beta(\frac{\nu-1}{2},\frac{\nu-1}{2})$ distribution. Since extensive incomplete-beta integral tables are available, (2.2) can be easily evaluated for any given values of $\bar{x},\bar{y}$ and s obtained from sample observations. Next, denoting

$$w = \frac{1}{2} - \frac{\bar{x}-\bar{y}}{2s\sqrt{\nu(4-(\frac{1}{n_1}+\frac{1}{n_2}))}}\ ,$$

(2.2) can be rewritten in the form

$$\hat{\Phi}(-\tfrac{\Delta}{2}) = \frac{1}{B(\frac{\nu-1}{2}\ \frac{\nu-1}{2})} \int_0^w [u(1-u)]^{(\nu-3)/2}du. \qquad (2.3)$$

3. Relative Efficiencies of MVUE and MLE of $\Phi(-\frac{\Delta}{2})$

First we derive formulas for the variance of MVUE, $\hat{\Phi}(-\Delta/2)$, and the MSE of MLE, $\Phi(-\hat{\Delta}/2)$, in forms suitable for numerical computations, and then present their values as well as relative efficiencies of both types of estimates for certain parametric values of Δ in each of the two cases (i) and (ii).

3.1. R.E. of the MVUE

Case (i). Noting that $v = \frac{\bar{x}-\bar{y}}{\sigma}$ is distributed normally with mean Δ and variance $(\frac{1}{n_1} + \frac{1}{n_2})$ and

$$\hat{\Phi}(-\tfrac{\Delta}{2}) = \text{Prob}\{Z \leq -\frac{v}{\sqrt{4-(\frac{1}{n_1}+\frac{1}{n_2})}} \mid v\},$$

we have

$$E[\hat{\Phi}(-\tfrac{\Delta}{2})]^2 = \int_{-\infty}^{\infty} \Phi^2(-\frac{v}{\sqrt{4-(\frac{1}{n_1}+\frac{1}{n_2})}})\ \phi(\frac{v-\Delta}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}})\ dv$$

$$= \Phi(-\tfrac{\Delta}{2}, -\tfrac{\Delta}{2}, \tfrac{1}{4}(\tfrac{1}{n_1}+\tfrac{1}{n_2})) \qquad (3.1)$$

where $\Phi(x,y;\rho)$ is the bivariate standard normal c.d.f. The result in (3.1) easily follows by a simple probability argument, e.g. see Zacks and Evens [9]. Hence

$$\text{Var}[\hat{\Phi}(-\tfrac{\Delta}{2})] = \Phi(-\tfrac{\Delta}{2}, -\tfrac{\Delta}{2}; \tfrac{1}{4}(\tfrac{1}{n_1}+\tfrac{1}{n_2})) - \Phi^2(-\tfrac{\Delta}{2}) \qquad (3.2)$$

and

$$\text{R.E.}[\hat{\Phi}(-\tfrac{\Delta}{2})] = (\tfrac{1}{n_1}+\tfrac{1}{n_2})\ \phi^2(-\tfrac{\Delta}{2})/4\ \text{Var}[\hat{\Phi}(-\tfrac{\Delta}{2})]\ . \qquad (3.3)$$

Case (ii). Letting $t = (\bar{x}-\bar{y})/s(\frac{1}{n_1}+\frac{1}{n_2})^{1/2}$, a non-central student t variate with ν d.f. and non-centrality parameter $\Delta/(\frac{1}{n_1}+\frac{1}{n_2})^{1/2}$, (2.2) can be written as

$$\hat{\Phi}(-\tfrac{\Delta}{2}) = \text{Prob}\{(2U-1) \leq -t\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}}/\sqrt{\nu[4-(\tfrac{1}{n_1}+\tfrac{1}{n_2})]} \mid t\}\ .$$

Furthermore, letting U_1 and U_2 be two independent beta variables, each distributed as $\beta(\frac{\nu-1}{2}, \frac{\nu-1}{2})$ and considering $W = \max(2U_1-1, 2U_2-1)$, it follows that the density function of W is

$$f(w) = \frac{1}{2^{\nu-3}B(\frac{\nu-1}{2},\frac{\nu-1}{2})}(1-w^2)^{\frac{\nu-3}{2}}\ I_{(1+w)/2}(\tfrac{\nu-1}{2}, \tfrac{\nu-1}{2}),\ -1 \leq w \leq 1$$

where

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x y^{a-1}(1-y)^{b-1}dy \ .$$

Now we can write

$$[\hat{\Phi}(-\tfrac{\Delta}{2})]^2 = \text{Prob}\ (W \leq -\frac{t\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}{\sqrt{\nu[4-(\frac{1}{n_1}+\frac{1}{n_2})]}} \mid t)$$

$$= F(-\frac{t\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}{\sqrt{\nu[4-(\frac{1}{n_1}+\frac{1}{n_2})]}} \mid t), \text{ (say)} \ .$$

Then

$$E[\hat{\Phi}(-\tfrac{\Delta}{2})]^2 = \int_{-\infty}^{\infty} F(-\frac{t\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}{\sqrt{\nu[4-(\frac{1}{n_1}+\frac{1}{n_2})]}})\ g(t;\delta,\nu)\ dt \qquad (3.4)$$

where $g(t;\delta,\nu)$ is a non-central student t density function with non-centrality parameter $\delta = \Delta/(\frac{1}{n_1}+\frac{1}{n_2})^{1/2}$ and ν d.f. For numerical integration of the right side in (3.4), we have considered the following forms for the functions $F(w)$ and $g(t;\delta,\nu)$:

$$F(w) = \frac{4}{(\nu-1)B^2(\frac{\nu-1}{2},\frac{\nu-1}{2})}\left[\int_0^{(1-w)/2} y^{\nu-2}(1-y)^{\nu-2}dy\right.$$

$$\left. + \sum_{k=0}^{\infty} \frac{B(\frac{\nu+1}{2},\ k+1)}{B(\nu-1,k+1)} \int_0^{(1+w)/2} y^{k+\nu-1}(1-y)^{\nu-2}dy\right]$$

where exact integration was obtained using recursive scheme, and

$$g(t,\delta,\nu) = \frac{1}{\sqrt{2\pi\nu}\ 2^{\nu/2}\Gamma(\frac{\nu}{2})} \int_0^{\infty} y^{(\nu-1)/2} \exp[-\tfrac{1}{2}\{(t\sqrt{\tfrac{y}{\nu}}-\delta)^2-y\}]dy$$

where the 32-point Gauss-Laguerre quadrature formula was used for numerical integrations. The final integration in (3.4) was done using the Rhomberg method of numerical integration. All of these evaluations were on double precision basis.

After finding the variance of $\hat{\Phi}(-\frac{\Delta}{2})$, we now obtain

$$R.E.[\hat{\Phi}(-\tfrac{\Delta}{2})] = (\frac{1}{n_1} + \frac{1}{n_2} + \frac{\Delta^2}{2(n_1+n_2-1)})\phi^2(-\tfrac{\Delta}{2})/4\ Var[\hat{\Phi}(-\tfrac{\Delta}{2})] \ . \tag{3.5}$$

3.2. R.E. of the MLE

For the MSE and Bias of the MLE, we need to find both $E[\Phi(-\hat{\Delta}/2)]$ and $E[\phi^2(-\hat{\Delta}/2]$. We will be using the following result of Ellison [5] in our discussion.

If X is a normal variate with mean μ and variance σ^2, then

$$E[\Phi(X)] = \Phi(\frac{\mu}{\sqrt{1+\sigma^2}}) \ . \tag{3.6}$$

Case (i): Since $-(\bar{x}-\bar{y})/2\sigma$ has the normal distribution with mean $-\Delta/2$ and variance $\frac{1}{4}(\frac{1}{n_1} + \frac{1}{n_2})$, it follows from (3.6) that

$$E[\Phi(-\tfrac{\hat{\Delta}}{2})] = \Phi(-\frac{\Delta}{2\sqrt{1+\frac{1}{4}(\frac{1}{n_1} + \frac{1}{n_2})}}).$$

Again, it can be easily shown that

$$E[\phi^2(-\tfrac{\hat{\Delta}}{2})] = \Phi(-\frac{\Delta}{2\sqrt{1+\frac{1}{4}(\frac{1}{n_1} + \frac{1}{n_2})}}, -\frac{\Delta}{2\sqrt{1+\frac{1}{4}(\frac{1}{n_1} + \frac{1}{n_2})}}, (1+\frac{4n_1n_2}{n_1+n_2})^{-1}) \tag{3.7}$$

Hence

$$MSE[\Phi(-\tfrac{\hat{\Delta}}{2})] = \Phi(\xi,\xi, (1+\frac{4n_1n_2}{n_1+n_2})^{-1} - 2\Phi(\xi)\Phi(-\tfrac{\Delta}{2}) + \phi^2(-\tfrac{\Delta}{2}) \tag{3.8}$$

and

$$\text{Bias}[\Phi(-\tfrac{\hat{\Delta}}{2})] = \Phi(\xi) - \Phi(-\tfrac{\Delta}{2}) \tag{3.9}$$

where

$$\xi = -\frac{\Delta}{2\sqrt{1+\frac{1}{4}(\frac{1}{n_1}+\frac{1}{n_2})}} \quad .$$

Next, the relative efficiency of $\Phi(-\frac{\hat{\Delta}}{2})$,

$$\text{R.E.}[\Phi(-\tfrac{\hat{\Delta}}{2})] = (\tfrac{1}{n_1}+\tfrac{1}{n_2})\,\phi^2(-\tfrac{\Delta}{2})/4\text{MSE}[\Phi(-\tfrac{\hat{\Delta}}{2})] \ . \tag{3.10}$$

Case (ii). Given s^2, conditionally $(\bar{x}-\bar{y})/2s$ has a normal distribution with mean $(\mu_1-\mu_2)/2s$ and variance $(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2/4s^2$. So by applying (3.6), we have

$$E[\Phi(-\tfrac{\hat{\Delta}}{2})] = E[E[\Phi(-\tfrac{\hat{\Delta}}{2})\,|s^2]\,]$$

$$= E[\Phi(-\Delta/\sqrt{\frac{4s^2}{\sigma^2}+(\frac{1}{n_1}+\frac{1}{n_2})}\,)] \ .$$

Since $Q = \nu s^2/\sigma^2$ has a χ^2 distribution with ν d.f.,

$$E[\Phi(-\tfrac{\hat{\Delta}}{2})] = E[\Phi(-\Delta/\sqrt{\frac{4Q}{\nu}+(\frac{1}{n_1}+\frac{1}{n_2})}\,)] \tag{3.11}$$

where the expectation on the right side is with respect to the random variable Q.

Next by drawing analogy with the previous result in (3.7), it can easily be shown that

$$E[\Phi^2(-\tfrac{\hat{\Delta}}{2})] = E[E[\Phi^2(-\tfrac{\hat{\Delta}}{2})\,|s^2]]$$

$$= E[\Phi(-\frac{\Delta}{\sqrt{\frac{4Q}{\nu}+(\frac{1}{n_1}+\frac{1}{n_2})}},\ -\frac{\Delta}{\sqrt{\frac{4Q}{\nu}+(\frac{1}{n_1}+\frac{1}{n_2})}},\ \frac{1}{1+\frac{4Q}{\nu}})]. \tag{3.12}$$

Accordingly,

$$\text{MSE}[\Phi(-\tfrac{\hat{\Delta}}{2})] = E[\Phi(\eta,\eta,\tfrac{1}{1+4Q/\nu})] - 2E[\Phi(\eta)]\Phi(-\tfrac{\Delta}{2}) + \Phi^2(-\tfrac{\Delta}{2}) \tag{3.13}$$

and

$$\text{Bias}[\Phi(-\tfrac{\hat{\Delta}}{2})] = E[\Phi(\eta)] - \Phi(-\tfrac{\Delta}{2}) \tag{3.14}$$

where

$$\eta = -\frac{\Delta}{\sqrt{\frac{4Q}{\nu} + (\frac{1}{n_1} + \frac{1}{n_2})}} \quad .$$

Once again,

$$\text{R.E.}[\Phi(-\tfrac{\hat{\Delta}}{2})] = (\frac{1}{n_1} + \frac{1}{n_2} + \frac{\Delta^2}{2(n_1+n_2-1)})\Phi^2(-\tfrac{\Delta}{2})/4\text{MSE}[\Phi(-\tfrac{\hat{\Delta}}{2})] \ . \tag{3.15}$$

Expressions in (3.1), (3.7) are in terms of the standard bivariate normal c.d.f. and can be easily evaluated using the method by Owen (1956), among others, for any given values of n_1, n_2 and Δ. Next, for the final numerical integration in (3.11) and (3.12), we employed the Gauss-Laguarre quadrature formula.

4. Numerical Results

Certain numerical results are presented in tables 1 and 2 considering $n_1=n_2=n$ and specifying values for n and Δ: n=5,10,15,30 and Δ = .5,1.0,1.5, 2.0,2.5,3.0 in table 1 for σ^2 known case and n=5,10 and Δ=.5,1.0,1.5,2.0, 2.5,3.0,3.5 in table 2 for σ^2 unknown case. (Due to limited computational facilities, we considered only two values for n in the latter case). The presented results are mainly designed to exemplify the comparison of the MVUE and the MLE in small sample case.

For the case of known σ^2 (figure 1), bias of the MLE is non-negative for all values of Δ, maximizing at Δ = 2.0. But when σ^2 is unknown (figure 2), bias is negative for Δ equal or less than 2.0 and positive for Δ greater than 2.0. However, in both cases, as one would expect, it decreases uniformly as the sample size n increases and is zero when Δ=0.

Further, interestingly enough, we observe that whether or not σ^2 is known makes a great difference in relative efficiences of the two types of estimates. The relative efficiency of MLE is higher than that of MVUE for smaller values of Δ (i.e. when probability of misclassification is higher) when σ^2 is known (Table 1). But reverse is the case when σ^2 is unknown (Table 2).

TABLE 1

Relative Efficiencies of MVUE and MLE of $\Phi(-\frac{\Delta}{2})$ for σ^2 Known Case

Δ	n	$\mathrm{Var}[\hat{\Phi}(-\frac{\Delta}{2})]$	$\mathrm{MSE}[\Phi(-\hat{\frac{\Delta}{2}})]$	$\mathrm{R.E.}[\hat{\Phi}(-\frac{\Delta}{2})]$	$\mathrm{R.E.}[\Phi(-\hat{\frac{\Delta}{2}})]$
.5	5	.01502	.01374	.9955	1.0881
	10	.00749	.00716	.9982	1.0044
	15	.00499	.00484	.9989	1.0297
	30	.00249	.00245	.9995	1.0149
1.0	5	.01256	.01172	.9867	1.0573
	10	.00624	.00603	.99361	1.0285
	15	.00415	.00406	.9958	1.0189
	30	.00207	.00205	.9980	1.0095
1.5	5	.00933	.00899	.9722	1.0082
	10	.00460	.00452	.9861	1.0026
	15	.00305	.00302	.9907	1.0014
	30	.00152	.00151	.9954	1.0005
2.0	5	.00615	.00620	.9523	.9438
	10	.00300	.00302	.9756	.9677
	15	.00198	.00200	.9836	.9774
	30	.00098	.00099	.9917	.9881
2.5	5	.00360	.00384	.9274	.8685
	10	.00173	.00180	.9623	.9249
	15	.00114	.00117	.9745	.9475
	30	.00056	.00057	.9871	.9725
3.0	5	.00186	.00212	.9014	.7907
	10	.00088	.00095	.9500	.8798
	15	.00058	.00061	.9653	.9144
	30	.00028	.00029	.9816	.9538

TABLE 2

Relative Efficiencies of MVUE and MLE of $\Phi(-\frac{\Delta}{2})$ for σ^2 Unknown Case

Δ	n	$Var[\hat{\Phi}(-\frac{\Delta}{2})]$	$MSE[\hat{\Phi}(-\frac{\Delta}{2})]$	$R.E.[\hat{\Phi}(-\frac{\Delta}{2})]$	$R.E.[\hat{\Phi}(-\frac{\Delta}{2})]$
.5	5	.01639	.01726	.9438	.8965
	10	.00780	.00814	.9900	.9485
1.0	5	.01491	.01494	.9467	.9448
	10	.00712	.00722	.9843	.9714
1.5	5	.01249	.01178	.9526	1.0106
	10	.00597	.00586	.9851	1.0027
2.0	5	.00951	.00849	.9576	1.0730
	10	.00454	.00432	.9845	1.0340
2.5	5	.00653	.00564	.9549	1.1052
	10	.00309	.00289	.9832	1.0514
3.0	5	.00403	.00346	.9374	1.0906
	10	.00188	.00175	.9747	1.0441
3.5	5	.00224	.00189	.8988	1.0615
	10	.00102	.00091	.9543	1.0717

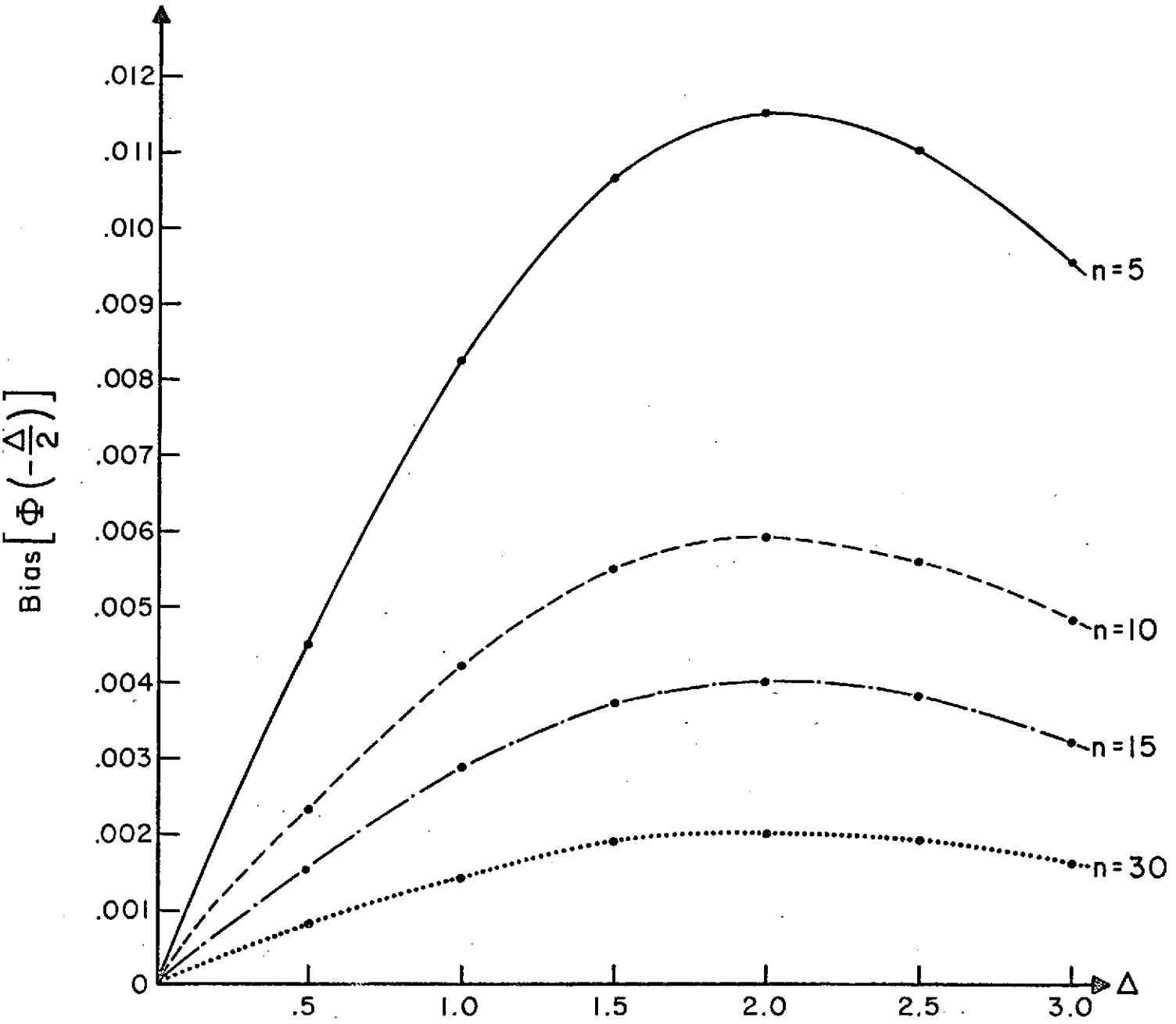


Figure 1: Bias of the MLE of the true probability of misclassification when σ^2 is known.

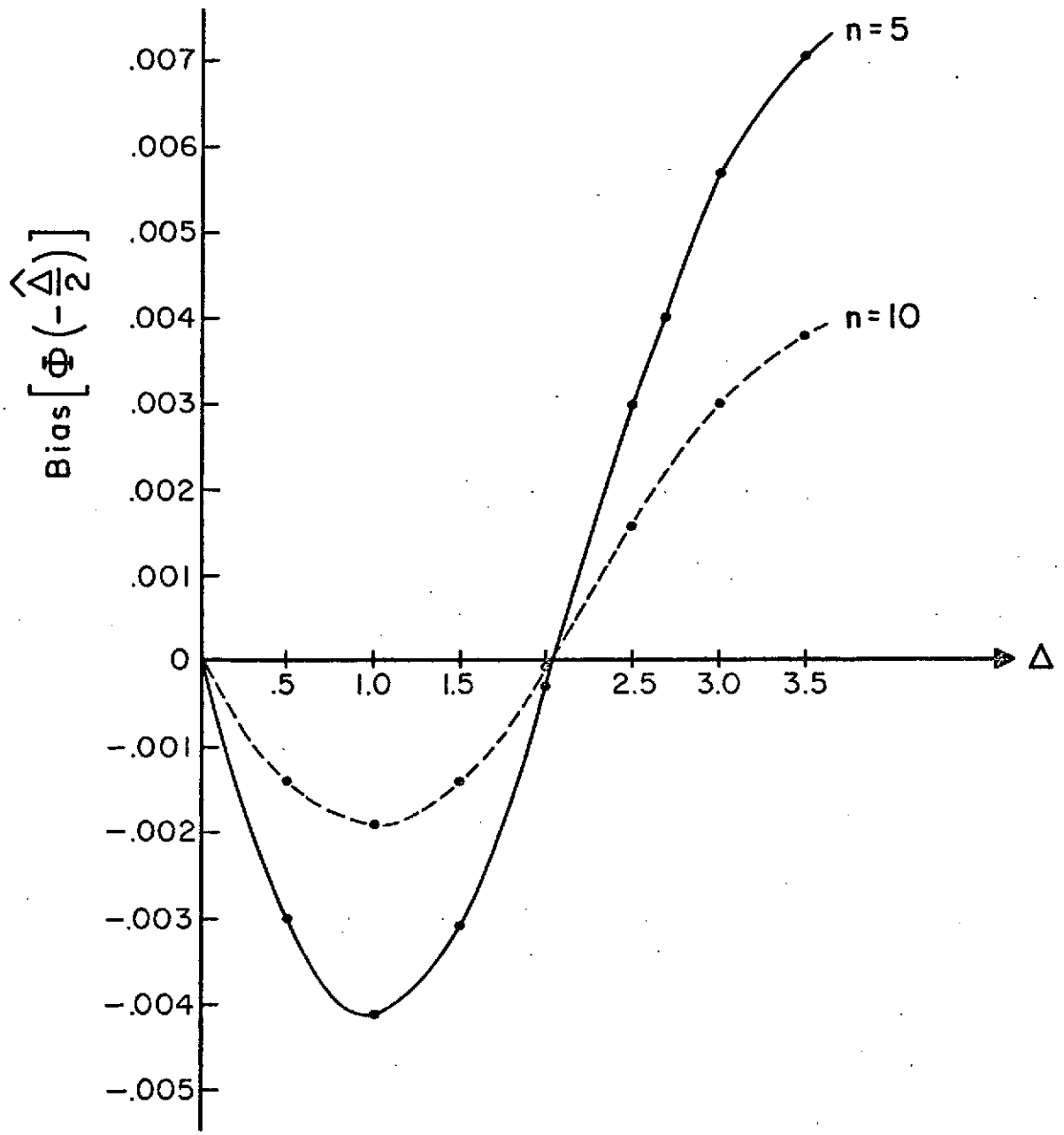


Figure 2: Bias of the MLE of the true probability of misclassification when σ^2 is unknown.

Acknowledgement

This research was supported by NASA, Johnson Space Center, Houston, Texas, Contract NAS9-13512.

The authors would like to thank Mr. Lynn Ziegler for his help in computations of tables 1 and 2.

References

[1] Abramowitz, M. and Stegun, I. A. Handbook of Mathematical Functions, National Bureau of Standards, Applied Mathematical Series 55, 1967.

[2] Chhikara, R. S. "Estimating proportions of objects under uncertainty" (Abstract). Institute of Mathematical Statistics Bulletin, 2, (May 1973), 98-99.

[3] Cochran, W. G. "Errors of Measurement in Statistics," Technometrics, 10 (November 1968), 637-66.

[4] Cramer, Harold. Mathematical Methods of Statistics, Princeton: Princeton University Press, 1946.

[5] Ellison, B. E. "Two theorems for inferences about the normal distributions with applications in acceptance sampling," Journal of American Statistical Association, 59 (March 1964), 89-96.

[6] Hills, M. "Allocation Rules and Their Error Rates," Journal of the Royal Statistical Society, Series B, 28 (1966), 1-32.

[7] Owen, D. B. "Tables for Computing Bivariate Normal Probabilities," Annals of Mathematical Statistics, 27 (1956), 1075-90.

[8] Sorum, M. "Estimating the expected probability of misclassification for a rule based on the linear discriminant function: Univariate normal case," Technometrics, 15 (May 1973), 329-39.

[9] Zacks, S. and Even, M. "The efficiencies of small samples of the maximum likelihood and best unbiased estimators of reliability functions," Journal of American Statistical Association, 61 (December 1966), 1033-51.

[10] Zacks, S. and Milton, R. "Mean square errors of the best unbiased and maximum likelihood estimators of tail probabilities in normal distributions," Journal of American Statistical Association, 66 (September 1971), 590-93.

A SIMULATION STUDY OF POPULATION CLASSIFICATION USING SELECTED VARIATES

by

Ernest R. Knezek,Jr. and Thomas L. Boullion

This research was supported in part by Johnson Space Center Contract #NAS9-13512

Summary

Considering the problem of estimating the optimum probability of misclassification under the maximum likelihood discriminant rule for univariate normal populations with common variance, we give its minimum variance unbiased estimate (MVUE) and derive the variance of the estimate for both cases of variance known and unknown. Next, we obtain the mean square error of the maximum likelihood estimate (MLE) and compare the two estimates by evaluating their relative efficiencies with respect to the Rao-Cramer lower bound for variances of any unbiased estimates. Also, bias of the MLE is investigated.

CHAPTER I

INTRODUCTION

The Bayesian Solution of the Discrimination Problem

Consider m populations $\Pi_1,\ldots,\Pi_m$ and suppose that each individual in the union of these populations possesses p common observable characteristics $C_1,\ldots,C_p$. Let the observed values of an individual be denoted by $X = (x_1,\ldots,x_p)^T$, where x_j denotes the observed values of C_j. The classical discriminant analysis problem consists of formulating a technique to divide the space of observations into m mutually exclusive and exhaustive regions $R_1,\ldots,R_m$ (hence classifying the individual in population Π_j if X falls in region R_j) in a manner such that the cost of misclassifying the individual is minimized.

There have been various techniques proposed for solving the problem, of which the Bayesian solution is optimal, in the sense that it minimizes the expected cost of misclassification. Anderson [1] states the Bayesian solution as follows:

> "If q_i is the a priori probability of drawing an observation from population Π_i with density $p_i(X)$ $(i = 1,\ldots,m)$ and if the cost of misclassifying an observation from Π_i as from Π_j is

$C(j|i)$, then the regions of classification, $R_1,\ldots,R_m$, that minimize the expected cost are defined by assigning X to R_k if:

$$\sum_{\substack{i=1\\ i\neq k}}^{m} q_i p_i(X) C(k|i) < \sum_{\substack{i=1\\ i\neq j}}^{m} q_i p_i(X) C(j|i) \qquad (1)$$

for $j = 1,\ldots,m, j \neq k$."

Estimation of Population Parameters

In actual situations the values $q_1,\ldots,q_m$, $C(i|j)$ for all $i, j = 1,\ldots,m$, and the probability densities $p_1(X)$, $p_2(X),\ldots,p_m(X)$ may not be known. One solution is to assume the populations are equally likely, $C(i|j) = 0$ for $i = j$ and $C(i|j)$ is constant for $i \neq j$, and the populations are normal, that is, when X is from population Π_i,

$$X \sim N(\mu_i, \Sigma_i) \qquad (2)$$

where μ_i is the mean vector and Σ_i is the covariance matrix of the ith population. In addition, if the parameters μ_i and Σ_i, $i = 1,\ldots,m$, are unknown and a random sample of size n_i (denoted $x_1^{(i)}, x_2^{(i)},\ldots, x_{n_i}^{(i)}$) can be obtained from the ith population, then μ_i and Σ_i can be estimated by the estimators $\bar{X}_i$ and $\hat{\Sigma}_i$ given by

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_j^{(i)} \qquad (3)$$

and

$$\hat{\Sigma}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_j^{(i)} - \bar{X}_i)(x_j^{(i)} - \bar{X}_i)^T \quad , \tag{4}$$

respectively [1]. One can then estimate

$$p_i(X) = p_i(X, \mu_i, \Sigma_i) \tag{5}$$

by

$$\hat{p}_i(X) = p_i(X; \bar{X}_i, \hat{\Sigma}_i) \tag{6}$$

or

$$\hat{p}_i(X) = (2\pi)^{-\frac{p}{2}} |\hat{\Sigma}_i|^{-\frac{1}{2}} \exp[-\tfrac{1}{2}(X - \bar{X}_i)^T \hat{\Sigma}_i^{-1} (X - \bar{X}_i)] \tag{7}$$

An individual with the observation vector $X = X_o$ will be classified as belonging to the jth population if

$$\max \hat{p}_i(X_o) = \hat{p}_j(X_o), \quad i = 1,2,\ldots,m. \tag{8}$$

CHAPTER II

DIMENSION REDUCTION AND VARIATE SELECTION

Misclassification as a Function of Separation

If p, the vector length, is large, the problem of classifying individuals becomes quite cumbersome and time consuming [7,9]. For example in the analysis of remote sensing data [10], when p = 12 the amount of computational time is immense. Thus it is desirable to reduce p while maintaining near optimal results. One technique is to pick an acceptable number of characteristics, say s, $s \leq p$, based on an examination of a measure of the separation between the populations, and proceed with Bayesian classification procedure using the s-variate marginal densities. Provided the separation is held constant or nearly so, it is feasible that fewer than p variates may be used for classification purposes without significant degradation of accuracy.

In the case where m = 2 and Π_1 and Π_2 are normal populations with unknown parameters μ_1, μ_2 and $\Sigma_1 = \Sigma_2 = \Sigma$, the classification procedure (8) can be replaced by an equivalent discriminant function [1]

$$V = X^T\hat{\Sigma}^{-1}(\bar{X}_1-\bar{X}_2)-\frac{1}{2}(\bar{X}_1+\bar{X}_2)^T\hat{\Sigma}^{-1}(\bar{X}_1-\bar{X}_2) \qquad (9)$$

where $\bar{X}_i$ is given by (3) and

$$\hat{\Sigma} = \frac{1}{n_1+n_2-2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_j^{(i)}-\bar{X}_i)(x_j^{(i)}-\bar{X}_i)^T. \qquad (10)$$

V is asymptotically distributed $N(-\alpha/2,\alpha)$ when $X \sim N(\mu_2,\Sigma)$ and $N(\alpha/2,\alpha)$ when $X \sim N(\mu_1,\Sigma)$, where α is the Mahalanobis distance between the populations given by

$$\alpha = (\mu_1-\mu_2)^T \Sigma^{-1}(\mu_1-\mu_2) \quad . \qquad (11)$$

Thus the probability of misclassification is asymptotically dependent on α and not on p. Hence the procedure to select $s < p$ in a manner so as to hold the separation, α, relatively constant, and then to classify on the basis of s variates, is reasonable.

Development of Divergence (A Measure of Separation)

For the two population case, $\Sigma_1 \neq \Sigma_2$, Kullback [5] uses the divergence between the populations as the measure of separation or difficulty of discriminating between the populations. Defining the logarithm of the likelihood ratio, $\log [p_1(X_o)/p_2(X_o)]$, to be the information in $X = X_o$ for discimination in favor of Π_1 against Π_2, the mean information for discrimination in favor of Π_1 against Π_2 is

$$I_{12} = \int p_1(X) \log \frac{p_1(X)}{p_2(X)} dX \; . \qquad (12)$$

The divergence between the populations, J, is defined by

$$J = I_{12} + I_{21} \quad . \tag{13}$$

In the case where Π_1 and Π_2 are s-variate normal populations,

$$p_i(X) = (2\pi)^{-\frac{s}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp[-\tfrac{1}{2}(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i)] \tag{14}$$

for i = 1,2. Then

$$\log \frac{p_1(X)}{p_2(X)} = \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} \operatorname{tr} \Sigma_1^{-1}(X-\mu_1)(X-\mu_1)^T + \frac{1}{2} \operatorname{tr} \Sigma_2^{-1}(X-\mu_2)(X-\mu_2)^T \quad , \tag{15}$$

from which we get

$$I_{12} = \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2} \operatorname{tr} \Sigma_1(\Sigma_2^{-1} - \Sigma_1^{-1}) + \frac{1}{2} \operatorname{tr} \Sigma_2^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^T \quad . \tag{16}$$

Similarly

$$I_{21} = \frac{1}{2} \log \frac{|\Sigma_1|}{|\Sigma_2|} + \frac{1}{2} \operatorname{tr} \Sigma_2(\Sigma_1^{-1} - \Sigma_2^{-1}) + \frac{1}{2} \operatorname{tr} \Sigma_1^{-1}(\mu_2-\mu_1)(\mu_2-\mu_1)^T \quad . \tag{17}$$

Thus

$$J = I_{12} + I_{21}$$

or

$$J = \frac{1}{2} \text{tr } (\Sigma_1-\Sigma_2)(\Sigma_2^{-1}-\Sigma_1^{-1}) + \frac{1}{2} \text{tr } (\Sigma_1^{-1}+\Sigma_2^{-1})(\mu_1-\mu_2)(\mu_1-\mu_2)^T . \tag{18}$$

Assuming equal covariance matrices, $\Sigma_1 = \Sigma_2 = \Sigma$,

$$J = \frac{1}{2} \text{tr } [2\Sigma^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^T] \tag{19}$$

or

$$J = (\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) \quad . \tag{20}$$

That is, if $\Sigma_1 = \Sigma_2 = \Sigma$,

$$J = \alpha \tag{21}$$

where α is the Mahalanobis distance between the populations.

It may be noted in (18) that the first term of the expression for divergence is due to the difference of the covariance matrices, and the second term is due to the difference of the means. In actual situations $\Sigma_1 \neq \Sigma_2$, however, one may choose to ignore the difference and compute α instead of J. Then a value to substitute for Σ in (11) or (20) is

$$\overline{\Sigma} = \frac{\Sigma_1+\Sigma_2}{2} \tag{22}$$

This assumption is made within this report, and the conclusions of Chapter IV indicate it is reasonable.

Sample Size Considerations

If the parameters of Π_1 and Π_2 are unknown but estimated on the basis of training samples of sizes n_1 and n_2, respectively, then the choice of s may be heavily dependent on the values n_1 and n_2. Consider 1) if $\Sigma_1 = \Sigma_2 = \Sigma$, $(n_1 + n_2)p$ measurements are used to estimate the $2p + (p^2 + p)/2$ distinct elements of $\bar{X}_1, \bar{X}_2$ and $\hat{\Sigma}$; and 2) if $\Sigma_1 \neq \Sigma_2$, $(n_1 + n_2)p$ measurements are used to estimate the $3p + p^2$ distinct elements of $\bar{X}_1, \bar{X}_2, \hat{\Sigma}_1$ and $\hat{\Sigma}_2$. In either case, the number of elements to be estimated increases as a function of p^2. This suggests that for small sample sizes the probability of classification may actually be improved by considering fewer variates. The results contained herein support this hypothesis.

CHAPTER III

SIMULATED POPULATION CLASSIFICATION

Classification Criteria

Anderson [1] discusses the classification when m = 2 and Π_1 and Π_2 are normal populations with equal covariance matrices. The classification procedure (8) can be replaced by an equivalent discriminant function

$$V = X^T\hat{\Sigma}^{-1}(\bar{X}_1-\bar{X}_2)-\frac{1}{2}(\bar{X}_1+\bar{X}_2)^T\hat{\Sigma}^{-1}(\bar{X}_1-\bar{X}_2) \qquad (23)$$

where $\bar{X}_i$ and $\hat{\Sigma}$ are given in (3) and (10), respectively. If $n_1 = n_2 = n$ then the distribution of V if X is from Π_1 is the same as the distribution of -V if X is from Π_2. Thus if $V \geq 0$ is the region of classification as Π_1, then the probability of misclassifying X when it is from Π_1 is equal to the probability of misclassifying X when it is from Π_2. Since V is asymptotically distributed $N(-\alpha/2,\alpha)$ when X is from Π_2, $P(1|2)$, the probability of classifying an observation from Π_2 in Π_1, is approximately

$$\int_0^\infty \frac{1}{\sqrt{2\pi\alpha}} e^{-(v+\alpha/2)^2/2\alpha} dv \quad . \qquad (24)$$

Similarly

$$P(2|1) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\alpha}} e^{-(v-\alpha/2)^2/2\alpha} dv \quad , \qquad (25)$$

and $P(2|1) = P(1|2)$.

In the event $\Sigma_1 \neq \Sigma_2$, $\hat{p}_1(X)$ and $\hat{p}_2(X)$ are estimated on the basis of training samples of sizes n_1 and n_2, respectively. X is classified in Π_2 if

$$\hat{p}_2(X) > \hat{p}_1(X) \tag{26}$$

and in Π_1 otherwise. Notice it is not necessary to compute $\hat{p}_1(X)$ and $\hat{p}_2(X)$ since (26) is equivalent to

$$|\hat{\Sigma}_2|^{\frac{1}{2}} \exp[-z_1/2] < |\hat{\Sigma}_1|^{\frac{1}{2}} \exp[-z_2/2] \tag{27}$$

where

$$z_i = (X-\overline{X}_i)^T \hat{\Sigma}_i^{-1} (X-\overline{X}_i) \quad , \tag{28}$$

$i = 1,2$. Taking the natural logarithm of (27) and multiplying through by 2 gives the classification rule: Classify X in Π_2 provided

$$\log (\det \hat{\Sigma}_1) - z_2 > \log (\det \hat{\Sigma}_2) - z_1 \tag{29}$$

and in Π_1 otherwise. This rule is equivalent to (8) and (26).

Monte Carlo Simulation Procedure

If the populations Π_1 and Π_2 are assumed to have equal covariances matrices, the probability $P(2|1)$ can be estimated from (25) provided the sample sizes n_1 and n_2 are large. For small (and moderately large) values of $n_1 = n_2 = n$, the probability $P(2|1)$ can be estimated by Monte Carlo methods. The process involves taking a sample of size n from each of

the two normal populations. These samples are used to compute $\bar{X}_1, \bar{X}_2$ and $\hat{\Sigma}$ according to (3) and (10). A sample population of 50 is generated from $\Pi_1 \sim N(\mu_1, \Sigma)$ and 50 values of V are calculated. The probability of classifying an individual from the Π_1 population as belonging to Π_2 is then estimated by

$$\hat{P}(2|1) = k/50 \qquad (30)$$

where k is the number of values of $V < 0$. The sequence [training samples - estimation - population - classification - $\hat{P}(2|1)$] is repeated 50 times, and the mean of the 50 values of $\hat{P}(2|1)$ is used as the estimate of $P(2|1)$. The results of Test Case 1 were obtained in this manner.

When the covariance matrices of Π_1 and Π_2 are not assumed to be equal, the procedure for estimating $P(2|1)$ is basically the same as the preceding discussion. Differences arise in the estimation of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ and in the classification procedure since V is no longer valid. $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are computed according to (4). The sample population of 50 is generated from $\Pi_1 \sim N(\mu_1, \Sigma_1)$ and the 50 members are classified as belonging to Π_1 or Π_2 according to (29). Then

$$\hat{P}(2|1) = k/50 \qquad (31)$$

where k is the number of times a member of the population is classified as belonging to Π_2. Again the sequence terminating in the computation of $\hat{P}(2|1)$ is repeated 50

times and the mean of the 50 values of $\hat{P}(2|1)$ is used as the final estimate of $P(2|1)$.

When $\Sigma_1 \neq \Sigma_2$, $\hat{P}(2|1) \neq \hat{P}(1|2)$. However, the estimation of $P(1|2)$ is accomplished by changing the order of the input of the parameters of the populations. The results of Test Case 2 were obtained in this manner.

CHAPTER IV

RESULTS AND CONCLUSIONS

Test Case 1: Feasibility

To demonstrate the feasibility of selecting a subset of the variates while holding the separation measure constant consider the populations Π_1 and Π_2 and the simulation results given in the following table:

TABLE 1. TEST CASE 1: FEASIBILITY

Number of Variates	$\mu_1=0$ μ_2	Max α	$\hat{P}(2\|1)$
1	(1)	1	.318
2	(1,1)	2	.256
3	(1,1,1)	3	.216
4	(1,1,1,1)	4	.171
5	(1,1,1,0,1)	4	.192
6	(1,1,1,0,1,0)	4	.195

$\Pi_1 \sim N(0,I)$ $\quad\quad$ $\mu_2 = (1,1,1,0,1,0)$

$\Pi_1 \sim N(\mu_2,I)$ $\quad\quad$ $n_1 = n_2 = 15$

Table 1 illustrates a hypothetical situation where the maximum separation, α, for the populations considered as 4-variate, 5-variate, and 6-variate populations is the same. Thus, if the training samples are large, an individual could be classified on the basis of four characteristics, namely C_1, C_2, C_3 and C_5, with no degradation of accuracy.

For small sample sizes, in particular $n_1 = n_2 = 15$, the simulation indicates that sample size considerations are significant, and better results are obtained using only 4 variates.

Test Case 2: Remote Sensing Data

To test the classification procedure using selected variates under meaningful conditions, actual data from a 12-channel sensor typical of those used in remote sensing of agricultural crops was obtained from NASA and used as population parameters for simulation purposes. Π_1 was a composite of soybean fields and is represented by μ_1 and Σ_1 as given in Table 2. Π_2 was a composite of corn fields and is represented by μ_2 and Σ_2 as given in Table 3. The divergence J for all combinations of 12 channels taken s at a time (s = 1,2,...,11) was computed by NASA. As many as 50 of the largest values of J were printed for each value of s and the values were included in the data package.

For purposes of comparing separation measures, α was computed [assuming $\Sigma = (\Sigma_1+\Sigma_2)/2$] for s = 1,3,6 and 12. The maximum and minimum α and the maximum J for the various values of s are plotted in Figure 1. The two distance measures are not readily comparable; however, the data does demonstrate the existence of combinations of fewer than 12 channels which yield almost the same separation as

Mean μ_1

169.56	174.76	193.24	192.64	169.11	166.80	190.56	171.49	185.49	173.79	160.94	181.18

Covariance Matrix Σ_1

19.08											
14.47	14.28										
9.01	7.52	5.91									
9.40	8.49	5.12	6.50								
16.77	14.59	9.00	9.86	19.23							
13.73	11.98	7.24	8.01	14.22	12.84						
7.97	7.01	4.68	4.90	8.44	7.15	5.31					
13.79	12.62	7.72	8.63	15.12	12.30	7.42	15.17				
10.43	10.02	6.00	7.00	12.08	9.82	6.14	11.47	10.88			
9.46	8.69	5.68	6.28	10.99	9.26	5.97	10.32	8.57	10.10		
-3.02	-3.48	-1.33	-2.19	-2.81	-1.10	-.20	-4.14	-4.02	-.84	23.59	
-4.73	-4.11	-2.73	-2.75	-4.24	-2.69	-1.67	-4.11	-3.23	-1.43	6.55	9.70

Table 2: MEAN AND COVARIANCE FOR SOYBEANS

Mean μ_2

172.35	178.58	196.25	196.02	175.20	171.77	194.28	180.16	193.56	183.46	157.35	178.62

Covariance Matrix Σ_2

23.92											
18.21	17.21										
11.57	9.33	7.25									
10.95	9.28	6.01	6.74								
23.40	19.28	12.65	12.59	28.43							
24.52	20.50	13.18	13.47	28.94	32.71						
13.29	11.06	7.52	7.61	15.95	17.52	10.84					
15.21	13.26	8.88	9.19	19.14	20.45	11.46	16.43				
8.93	8.12	5.41	5.76	11.49	12.11	6.97	9.92	7.76			
9.14	8.16	6.26	6.65	14.15	15.62	9.31	12.40	8.32	13.89		
19.88	16.31	11.27	11.92	27.21	32.89	18.61	18.00	8.77	17.48	63.35	
8.79	7.52	5.17	5.72	13.18	15.85	9.01	9.57	4.94	9.32	26.18	18.21

Table 3: MEAN AND COVARIANCE FOR CORN

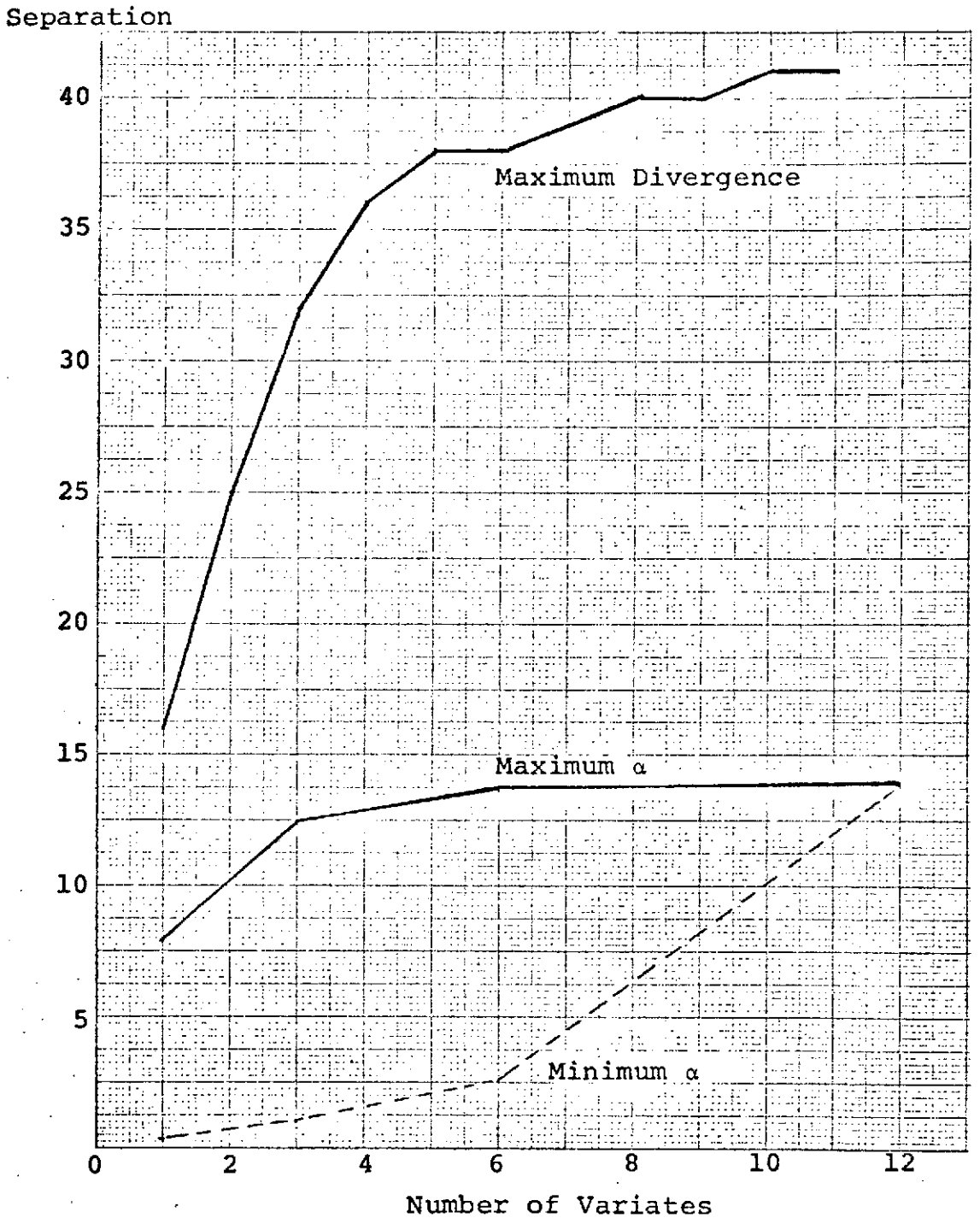


FIGURE 1. SEPARATION vs. NUMBER OF VARIATES

the full complement of channels. In particular, regardless of the choice of separation measure, very little increase in maximum separation is obtained by considering more than 5 channels.

To further study the equivalence of J and α as the criteria for the selection of variates, Table 4 was constructed. Table 4 contains some of the values obtained for J and α for various subsets of variates. For a single variate, the largest three values of J arise from considering the same three variates or channels that yielded the largest values of α, and in the same order.

TABLE 4. SEPARATION AS A SELECTOR OF SUBSETS

Subset of Channels	Values of α	Values of J
(10)	7.8*	16**
(9)	7.0	15
(8)	4.8	10
(6,10,12)	12.0	32**
(6, 9,10)	12.4*	31
(1,10,12)	12.3	30
(4, 6, 9,10,11,12)	13.5	38**
(4, 6, 8, 9,10,12)	13.6	38
(4, 6, 7, 9,10,12)	13.5	38
(1, 4, 6, 9,10,12)	13.6	38
(1, 4, 8, 9,10,12)	13.7*	

* largest α for fixed number of channels
** largest J for fixed number of channels

For the three variate case, the best three subsets by J measure were the best three subsets by α measure, though in different order. In the case of 6 variates, from the

924 possible subsets of channels, the best three by J measure were not among the best three by α measure; however, the best three by J did yield values of α significantly close to the maximum α.

The conclusion from Figure 1 and Table 4 is that either of the separation measures between populations, divergence and Mahalanobis distance, is a suitable selection criteria for the desired subset of variates.

Figures 2, 3, and 4 were constructed from simulation data based on training sample sizes of 15. They support the theory that $P(2|1)$ is a decreasing function of α. In most instances $\hat{P}(2|1)$ decreases as the distance measure increases. The relation is consistent with what could be expected from the asymptotic theory. Figure 2 is not too coherent because there are exactly 12 data points to consider. It does, however, indicate the trend. Figure 3 is very consistent in support of the relation. Figure 4 shows considerable scattering. If one looks ahead to Figure 6, the reason for the scatter in Figure 4 is evident. The sample size was too small. A training sample size of 30 would probably have produced a more consistent figure.

Figures 5 and 6 are the significant figures of this report. They represent the same data presented from two different viewpoints. Figure 5 shows that for small fixed training sample sizes, there is a subset of fewer than 12

variates which yield the minimum probability of misclassification. Figure 6 holds p constant and varies the training sample size. This figure shows that, for the populations considered, if one is constrained to $n \leq 11$, one might just as well look at a single variate. For $11 < n < 30$, 3-channel data is adequate, and n must be somewhere in the neighborhood of 100 before 12-variate data can be justified over 6-variate data.

The conclusion of this report is that for small sample sizes, a subset of the variates can be chosen so as to yield better classification results than the full complement of variates, and the selection of variates can be accomplished by considering the separation between the populations.

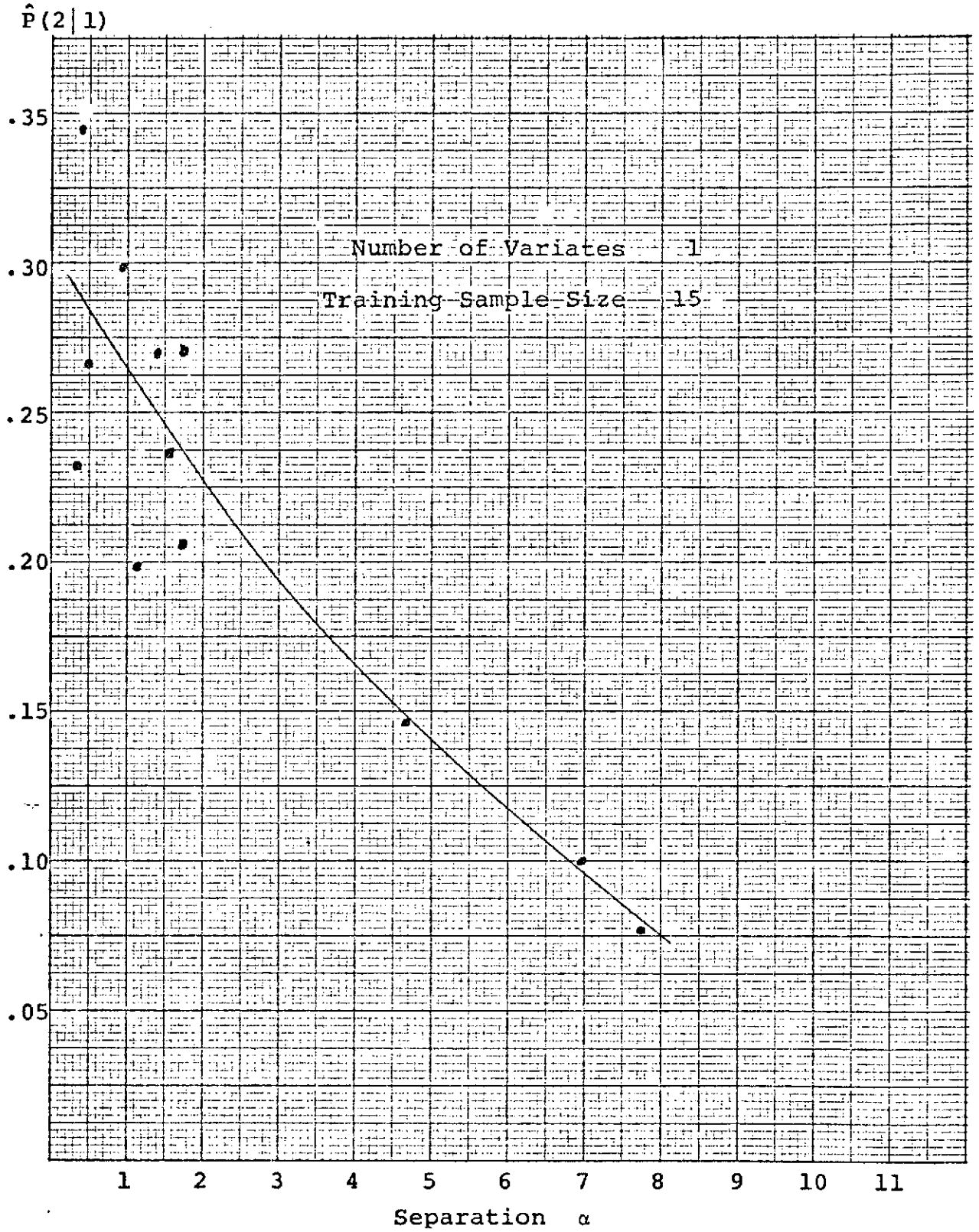


FIGURE 2. $\hat{P}(2|1)$ vs. α - SINGLE VARIATE

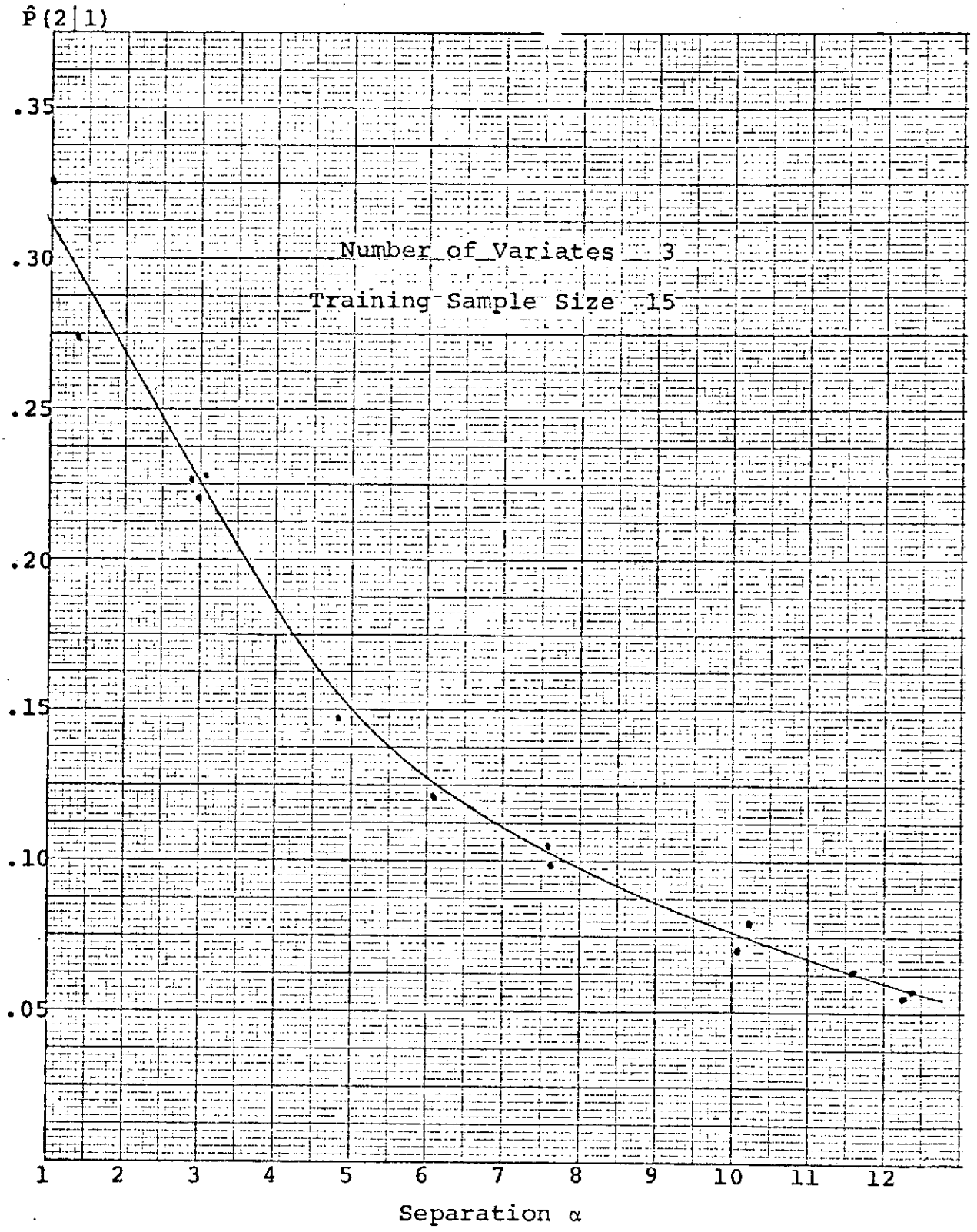


FIGURE 3. $\hat{P}(2|1)$ vs. α - THREE VARIATES

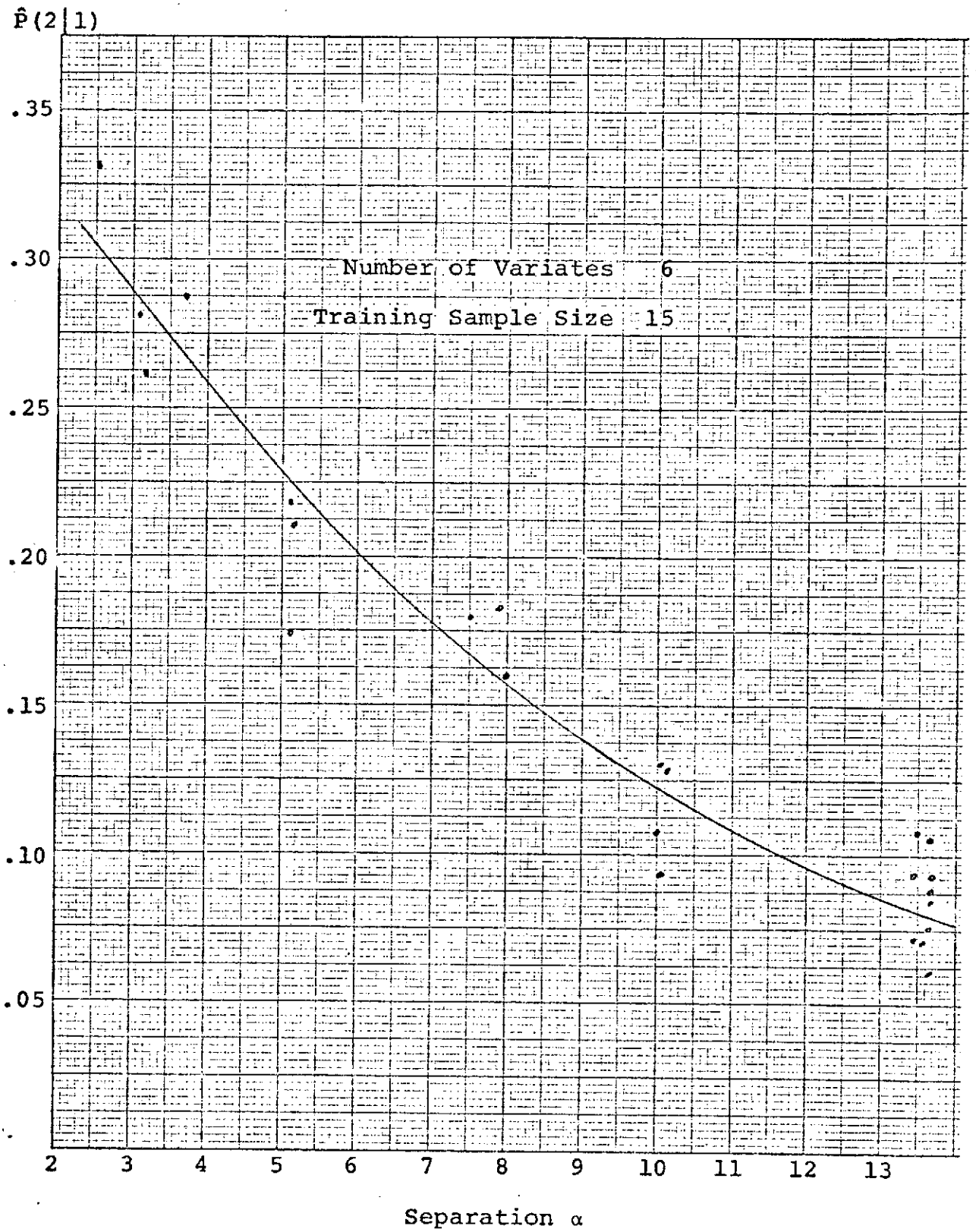


FIGURE 4. $\hat{P}(2|1)$ vs. α - SIX VARIATES

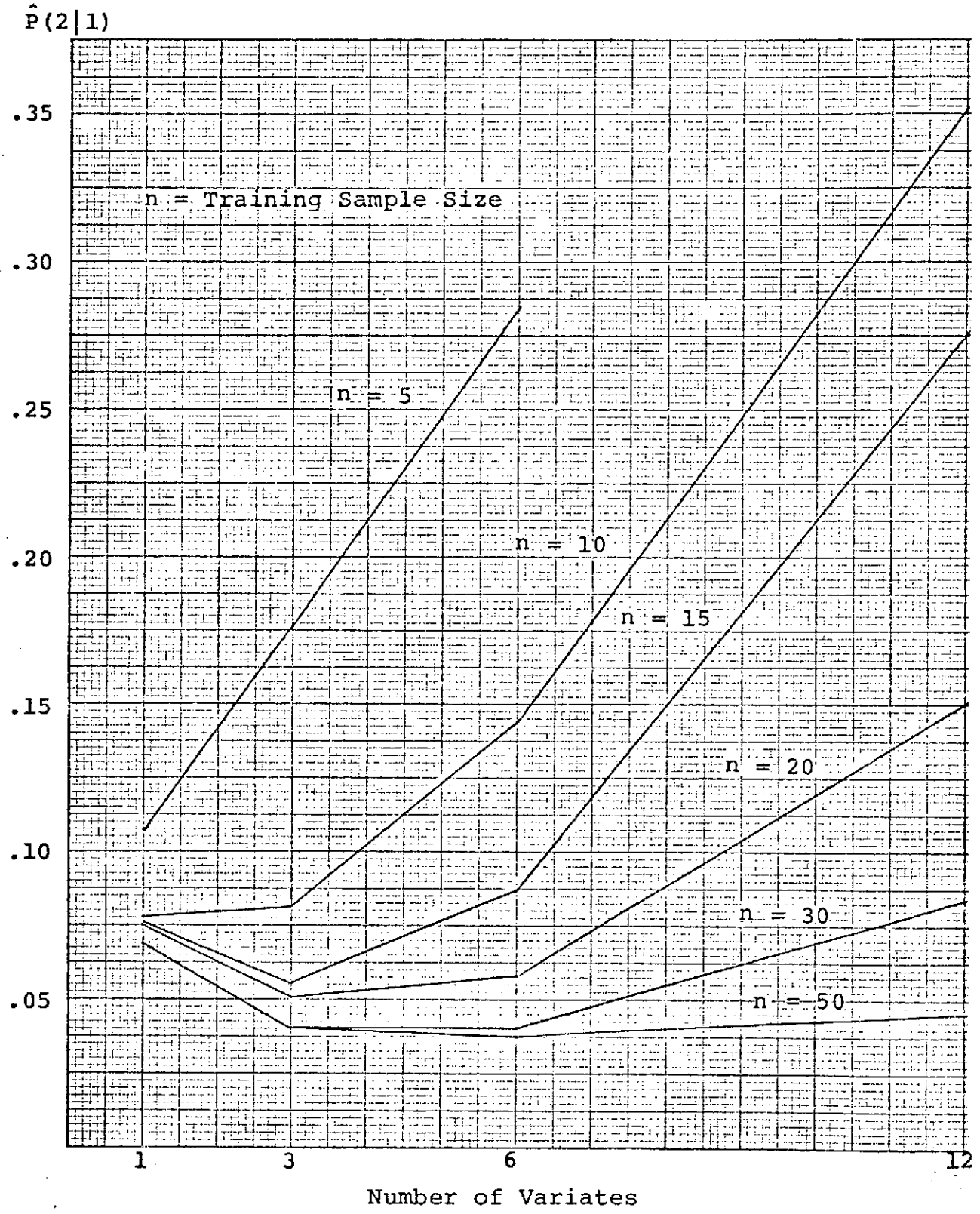


FIGURE 5. $\hat{P}(2|1)$ vs. NUMBER OF VARIATES FOR FIXED SAMPLE SIZES

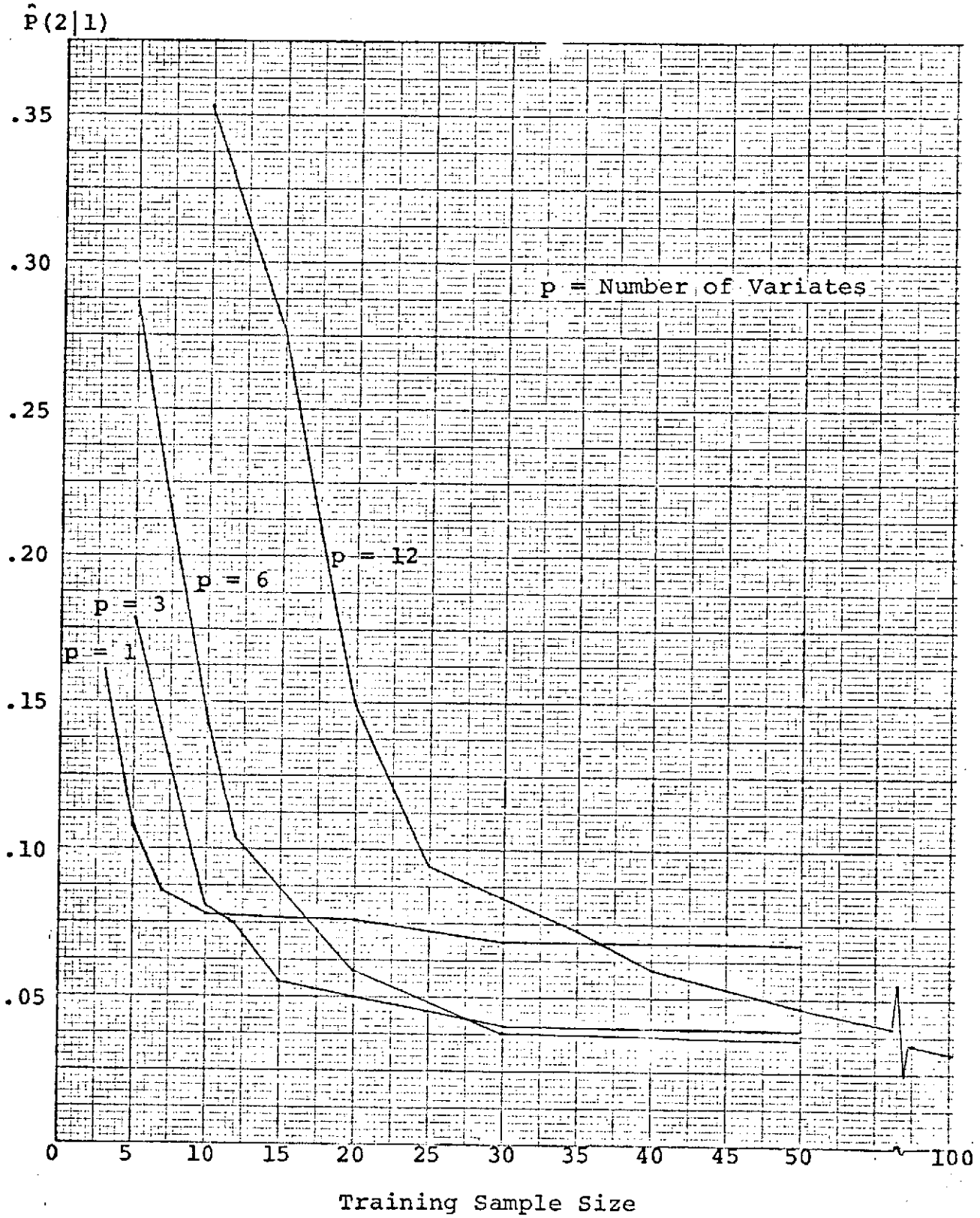


FIGURE 6. $\hat{P}(2|1)$ vs. TRAINING SAMPLE SIZE FOR FIXED VARIATES

BIBLIOGRAPHY

[1] Anderson, T.W., An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc., New York (1958).

[2] Cooley, W.W. and Lohnes, P.R., Multivariate Data Analysis, John Wiley & Sons, Inc., New York (1971).

[3] Duran, B.S., Gray, H.L., Tubbs, J. and Boullion, T.L., "On Estimating the Probability of Misclassification," Unpublished paper, Mathematics Department, Texas Tech University, Lubbock, Texas.

[4] Graybill, F.A., An Introduction to Linear Statistical Models, Volume 1, McGraw-Hill Book Company, Inc., New York (1961).

[5] Kullback, S., Information Theory and Statistics, John Wiley & Sons, Inc., New York (1959).

[6] Newman, T.G. and Odell, P.L., The Generation of Random Variates, Griffin's Statistical Monographs and Courses, No. 29, Hafner Publishing Company, New York (1971).

[7] Payne, M.P., "A Numerical Simulation of Wilk's Scatter Technique for Dimension Reduction in Statistical Discriminant Problems," A Thesis (M.S.) in Mathematics, Texas Tech University, Lubbock (1971).

[8] Perlis, S., Theory of Matrices, Addison-Wesley Publishing Company, Inc., Reading, Massachusetts (1952).

[9] Tubbs, J., Duran, B.S., Boullion, T.L. and Odell, P.L., "An Empirical Study of Classification by Thresholding," Unpublished paper, Mathematics Department, Texas Tech University, Lubbock, Texas.

[10] Remote Sensing of Earth Resources, NASA SP 7036, A Literature Survey with Indexes, September 1970.

ACREAGE ESTIMATES FOR CROPS USING REMOTE SENSING TECHNIQUES: CLASSIFICATION ERROR MATRIX KNOWN

R. S. Chhikara

and

P. L. Odell

The University of Texas at Dallas

*This research is supported by Johnson Space Center Contract #NAS9-13512.

ACREAGE ESTIMATES FOR CROPS USING REMOTE SENSING TECHNIQUES: CLASSIFICATION ERROR MATRIX KNOWN

1. Introduction

Remote sensing technology has shown a great potential for data collection for the earth resources in any given geographic region. As a result of this, it may be feasible to assess various earth resources and, thereby, answers to some of the important yet previously unsolvable problems may be available. At present we address ourselves to the problem of estimating crop sizes, i.e., amount of acreage under crops, in a large area using remote sensing technique. Even though it may not be difficult to obtain full data over an area by using remote sensors, crop acreage estimation on the basis of complete enumeration of data points in the area may not be feasible both from technical and economical viewpoints. As such it would be desirable to derive estimates based upon sampled data acquired by a suitable sampling process.

In this report we discuss a sampling scheme providing crop acreage estimates in a given geographic region, and investigate the precision of these estimates and the effect of misclassification on the estimates. Since observations are obtained on individual pixels, we consider the frame made of the collection of all pixels in the region of our interest. As to the actual layout for various crops, two cases arise. One is that no information is available and the other is that some a priori information on the physical situation of crops exists. For example, if the area of interest is small like a county or certain areas covering Western Oklahoma, it is

possible to have some prior information for the location of different crops. In our discussion we consider each of these cases and point out the difference that exists in precisions of two estimates.

2. Formulation of the Problem

Let $C_1, C_2, \ldots, C_m$ be m different crops and $A_1, A_2, \ldots, A_m$ be their corresponding acreage areas in a region. Assume the whole region consists of N total number of pixels. Since a pixel on the basis of observation taken by a remote sensor is subject to uncertainity in its correct identification, let $P(i|j)$ denote the probability of misclassifying a pixel from crop C_j into crop C_i, and $P(i|i)$ denote the probability of correctly classifying a pixel into crop C_i. Accordingly, the expected acreages associated with C_i on the basis of remote sensing data is

$$E_i = \sum_{j=1}^{m} A_j \, P(i|j) \ , \tag{2.1}$$

$i=1,2,\ldots,m$.

Given the total number of pixels and a pixel size, it is sufficient to consider proportions of pixels associated with different crops in the region. Let $p_1, p_2, \ldots, p_m$ be the proportions of pixels for $C_1, C_2, \ldots, C_m$, respectively. Then the expected proportions of pixels likely to be identified in C_i, $i=1,2,\ldots,m$, are

$$e_i = \sum_{j=1}^{m} p_j \, P(i|j) \ . \tag{2.2}$$

e_i coincides with $p_i, i=1,2,\ldots,m$, when every pixel is correctly classified. Since it is almost impossible to expect so in remote sensing, in our discussion we first consider estimation of e_i, $i=1,2,\ldots,m$, and then seek estimates

for p_i, i=1,2,...,m, assuming that $P(i|j) > 0$ at least for one j different from i. This will be done considering two cases: (a) $P(j|i)$'s are known and (b) $P(i|j)$'s are unknown.

3. The Sampling Procedure

As we mentioned earlier in Section 1, we consider the following two cases:

(i) Crop boundaries (i.e. the physical layout for crops) are unknown.

(ii) Crop boundaries are enhanced and so known.

(i) In this case, a sampling procedure which appears to be useful is a two-stage sampling scheme. At the first stage a specified number of flightlines, each of the same size, are randomly selected, and at the second stage an equal number of units (pixels) are randomly selected from each of the selected flightlines. However, in order for this scheme to be useful, it may require a fairly large number of flightlines to be selected. Otherwise, a simple random sampling with single-stage, though more difficult to execute, may produce estimates with smaller variance than the two-stage random sampling scheme.

(ii) When the physical layout for the crops is known to a certain degree, consider the region made of subregions, each consisting of flightlines covering area as homogeneous as possible. This could allow more than one crop in a subregion, and such subdivision should be restricted to a minimum possible number of subregions. For an illustration, see Figure 1 given at the end.

For a sampling procedure, we suggest a three-stage sampling scheme carried out in each subregion. At the first stage flightlines are randomly selected proportional to the size of a subregion; at the second stage blocks (or segments) are randomly selected within a flightline proportional to the size of a crop in the flightline; and at the third stage units within blocks are randomly selected in equal numbers. The number of blocks selected in each of the sampled flightlines is the same. Here by block we mean a sampling unit of specific size within a flightline. It may be pointed out that careful consideration should be given in specifying the size of a block. Neither should it be too small for any practical use, nor should it be too large to make any distinction for the crops when sampled. This way at the first two stages the sampling is proportional to the size of the underlying strata (i.e., subregions or crops, whichever may be the case) and at the last stage it is a simple random sampling. Though one may expect accumulation in sampling error due to three stages of sampling, it should lead to a smaller variance of the estimate for e_i $(i=1,2,\ldots,m)$ compared to a single-stage or double-stage sampling scheme.

4. Expected Proportions Estimates

(i) Let N denote the total number of flightlines for the region and let n flightlines out of N be randomly selected. Suppose each flightline consists of R units from which r units are randomly sampled for each of the sampled flightlines. In the sample, let m_{it} denote the number of units from tth selected flightline classified into C_i. Now denoting

$$\hat{e}_{it} = \frac{m_{it}}{r}, \qquad t = 1,2,\ldots,n,$$

an unbiased estimate of e_i is given by

$$\hat{e}_i = \frac{1}{nr} \sum_{t=1}^{n} m_{it}, \qquad i=1,2,\ldots,m \ . \tag{4.1}$$

Since the total size of units, NR, is expected to be large in relation to the sample size, nr, the random vector $\hat{e} = (\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_m)^T$ has a multinomial distribution, usually a satisfactory approximation in such a situation. So it follows that the variance of $\hat{e}_i$,

$$V(\hat{e}_i) = (1 - \frac{n}{N}) \frac{1}{n(N-1)} \sum_{t=1}^{N} (e_{it} - e_i)^2 + (1 - \frac{r}{R}) \frac{1}{nr} \frac{R}{N(R-1)} \sum_{t=1}^{N} e_{it}(1 - e_{it}) \tag{4.2}$$

and an unbiased estimate of $V(\hat{e}_i)$ is given by

$$\hat{V}(\hat{e}_i) = (1 - \frac{n}{N}) \frac{1}{n(n-1)} \sum_{t=1}^{n} (\hat{e}_{it} - \hat{e}_i)^2 + (1 - \frac{r}{R}) \frac{1}{Nn(r-1)} \sum_{t=1}^{n} \hat{e}_{it} (1 - \hat{e}_{it}) \tag{4.3}$$

where e_{it} denotes the expected proportion of C_i in tth flight-line and

$e_i = \sum_{t=1}^{N} e_{it}/N$, the same as defined in (2.2), (Cochran, 1963).

(ii) Again, let N denote the total number of flightlines, each consisting of R units. Suppose the region is divided into k subregions, $R_1, R_2, \ldots, R_k$, and $N_1, N_2, \ldots, N_k$ are the corresponding number of flightlines for $R_1, R_2, \ldots, R_k$ such that

$$N = \sum_{j=1}^{k} N_j \ .$$

Considering a sample of n flightlines out of N, let n_j flightlines be randomly selected from N_j in R_j such that $n_j = n(N_j/N)$, j=1,2,...,k, and

$$n = \sum_{j=1}^{k} n_j \ .$$

For subregion R_j, let B_{jit} be the number of blocks covering ith crop in the tth sampled flightlines; from this let b_{jit} blocks be randomly samples such that

$$B = \sum_{i=1}^{m} B_{jit} \ , \qquad b = \sum_{i=1}^{m} b_{jit}$$

and

$$b_{jit} = \frac{b}{B} B_{jit} \ ,$$

$$t = 1,2,\ldots,n_j,$$

$$i = 1,2,\ldots,m \ .$$

Considering block size L, suppose l units are randomly selected for each sampled block. Thus the sample size for subregion R_j is $n_j bl$ and the complete sample consists of nbl units.

Let u_{jit} denote the number of sample units being classified in C_i for tth flightline selected from R_j. Then

$$\hat{e}_{ji} = \frac{1}{n_j bl} \sum_{t=1}^{n_i} u_{jit} \tag{4.4}$$

is an estimate of e_{ji}, the expected proportion of units in C_i for subregion R_j, i=1,2,...,m and j=1,2,...,k. Thus

$$\hat{e}_i = \frac{1}{N} \sum_{j=1}^{k} N_j \hat{e}_{ji}, \qquad i=1,2,\ldots,m. \tag{4.5}$$

Next,

$$V(\hat{e}_i) = \frac{1}{N^2} \sum_{j=1}^{k} N_j V(\hat{e}_{ji})$$

where

$$V(\hat{e}_{ji}) = \frac{N}{n}\left(1-\frac{n}{N}\right)\cdot\frac{e_{ji}(1-e_{ji})}{N_j-1},$$

assuming the same variance in each flightline of R_j for a fixed C_i. Thus

$$V(\hat{e}_i) = \frac{1}{nN}\left(1-\frac{n}{N}\right)\sum_{j=1}^{k}\frac{N_j^2}{N_j-1}\,e_{ji}\,(1-e_{ji}), \qquad (4.6)$$

$$i = 1,2,\ldots,m.$$

5. Actual Proportions Estimates

Denoting

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix}, \qquad p = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{bmatrix}, \quad \text{and}$$

$$Q = \begin{bmatrix} P(1|1) & P(1|2) & \cdots & P(1|m) \\ P(2|1) & P(2|2) & \cdots & P(2|m) \\ \vdots & \vdots & \cdots & \vdots \\ P(m|1) & P(m|2) & \cdots & P(m|m) \end{bmatrix},$$

(2.2) can be written as

$$e = Qp. \qquad (5.1)$$

Observe that Q may be a singular matrix. In view of this the equations in (5.1) may or may not have a solution for p for specified e, Q. In case any solution exists, it is given by

$$p = Q^I e + p_0 \tag{5.2}$$

where Q^I is a generalized inverse of Q matrix satisfying

$$QQ^IQ = Q \tag{5.3}$$

and p_0 satisfying $Qp_0 = 0$.

In the second case when no solution exists, we look for the vector p that minimizes the squared length of e-Qp for a given e, Q. Any such vector is given by

$$p = Q^S e + p_o \tag{5.4}$$

where Q^S is a right pseudo inverse of Q matrix satisfying

$$(QQ^S)^T = QQ^S \tag{5.5}$$

and p_0 satisfying $Qp_0 = 0$.

Since more than one p_0 may satisfy the condition $Qp_0 = 0$, in either case, a unique solution for p may not exist. Accordingly, either one would have a complete set of solutions p given by (5.2) or a complete set of vectors p which are a "best-fit" when obtained from (5.4). The techniques for finding Q^I, Q^S and p_0 are well known (Boullion and Odell, 1972). Henceforth, by a solution we mean in either sense and will denote it by

$$p = Q^G e \ , \tag{5.6}$$

assuming $p_0 = 0$ without loss of generality.

5.1. Q known

(i) Taking the estimate $\hat{e}$ given by (4.1) and the known value of Q, it follows from the above discussion that an estimate of p is given by

$$\hat{p} = Q^G \hat{e} \ . \tag{5.7}$$

Denoting the (i,j) the element of Q^G by q^G_{ij}, we have

$$\hat{p}_i = \sum_{j=1}^{m} q_{ij}^{G}\, \hat{e}_j , \qquad i=1,2,\ldots,m. \tag{5.8}$$

These are unbiased estimates of p_i, $i=1,2,\ldots,m$. For the variance,

$$V(\hat{p}_i) = \sum_{j=1}^{m} (q_{ij}^{G})^2\, V(\hat{e}_j) + \sum_{j=1}^{m} \sum_{\substack{j'=1 \\ j\neq j'}}^{m} q_{ij}^{G}\, q_{ij'}^{G}\, \mathrm{Cov}(\hat{e}_j,\hat{e}_{j'}) \tag{5.9}$$

where $V(\hat{e}_j)$ is given in (4.2) and

$$\mathrm{Cov}(\hat{e}_j,\hat{e}_{j'}) = \left(1-\frac{n}{N}\right)\frac{1}{n(N-1)} \sum_{t=1}^{N} (e_{jt}-e_j)(e_{j't}-e_{j'})$$

$$- \left(1-\frac{r}{R}\right)\frac{1}{Nr}\,\frac{R}{N(R-1)} \sum_{t=1}^{N} e_{jt}e_{j't}, \quad (j\neq j')$$

Next, an unbiased estimate of $V(\hat{e}_j)$ is given in (4.3). Similarly an unbiased estimate of $\mathrm{Cov}\,(e_j,e_{j'})$ can be shown as

$$\mathrm{Cov}(\hat{e}_j,\hat{e}_{j'}) = \left(1-\frac{n}{N}\right)\frac{1}{n(n-1)} \sum_{t=1}^{n} (\hat{e}_{jt}-\hat{e}_j)(\hat{e}_{j't}-\hat{e}_{j'})$$

$$- \left(1-\frac{r}{R}\right)\frac{1}{Nr(r-1)} \sum_{t=1}^{n} \hat{e}_{jt}\hat{e}_{j't}, \quad (j\neq j') .$$

Replacing the unknown quantities in (5.9) by their estimates, one thus obtains an unbiased estimate of $V(\hat{p}_i)$, $i=1,2,\ldots,m$.

(ii) In this case using the estimate $\hat{e}$ given in (4.5) we obtain

$$\hat{p} = Q^{G}\hat{e} \tag{5.10}$$

or

$$\hat{p}_i = \sum_{j=1}^{m} q^G_{ij}\, \hat{e}_j, \qquad i=1,2,\ldots,m . \tag{5.11}$$

Once again,

$$V(\hat{p}_i) = \sum_{j=1}^{m} (q^G_{ij})^2\, V(\hat{e}_j) + \sum_{\substack{j=1 \\ j\neq j'}}^{m} \sum_{j'=1}^{m} q^G_{ij} q^G_{ij'} \mathrm{Cov}(\hat{e}_i, \hat{e}_{j'}) \tag{5.12}$$

where $V(\hat{e}_j)$ is given in (4.6) and

$$\mathrm{Cov}(\hat{e}_j, \hat{e}_{j'}) = -\frac{1}{nN}\left(1-\frac{n}{N}\right) \sum_{i=1}^{k} \frac{N_j}{N_j - 1}\, e_{ij} e_{ij'} .$$

Now replacing $V(\hat{e}_j)$ and $\mathrm{Cov}(\hat{e}_j, \hat{e}_{j'})$ by their estimates, we can get an estimate of $V(\hat{e}_j)$, $i=1,2,\ldots,m$.

5.2. Q unknown

In this case one also needs to estimate $P(i|j)$, i and $j=1,2,\ldots,m$. An obvious way to do so is to ascertain the ground truth for a certain number of additional observations taken independently of those used for estimating e_i's and utilize these to estimate $P(i|j)$'s. In view of this $\hat{e}$ and $\hat{Q}$ will be stochastically independent and an estimate of p is given by

$$\hat{p} = \hat{Q}^G \hat{e} \tag{5.13}$$

or

$$\hat{p}_i = \sum_{j=1}^{m} \hat{q}^G_{ij}\, \hat{e}_j , \qquad i=1,2,\ldots,m . \tag{5.14}$$

where $\hat{q}^G_{ij}$ is the (i,j)th element of $\hat{Q}^G$.

For a given $\hat{Q}^G$, the conditional variance of $\hat{p}_i$ is given by (5.9) in case (i) and (5.12) in case (ii) with q_{ij}^G's replaced by $\hat{q}_{ij}^G$'s. For the unconditional variance, the expression will involve the variances and covariances for the elements of $\hat{Q}^G$. In general it will be difficult to express the unconditional variance explicitly.

References

Cochran, W. Sampling Techniques, 2nd Edition, John Wiley and Sons, Inc., 1963.

Boullion, T. L. and Odell, P. L. Generalized Inverse Matrices, John Wiley and Sons, 1971.

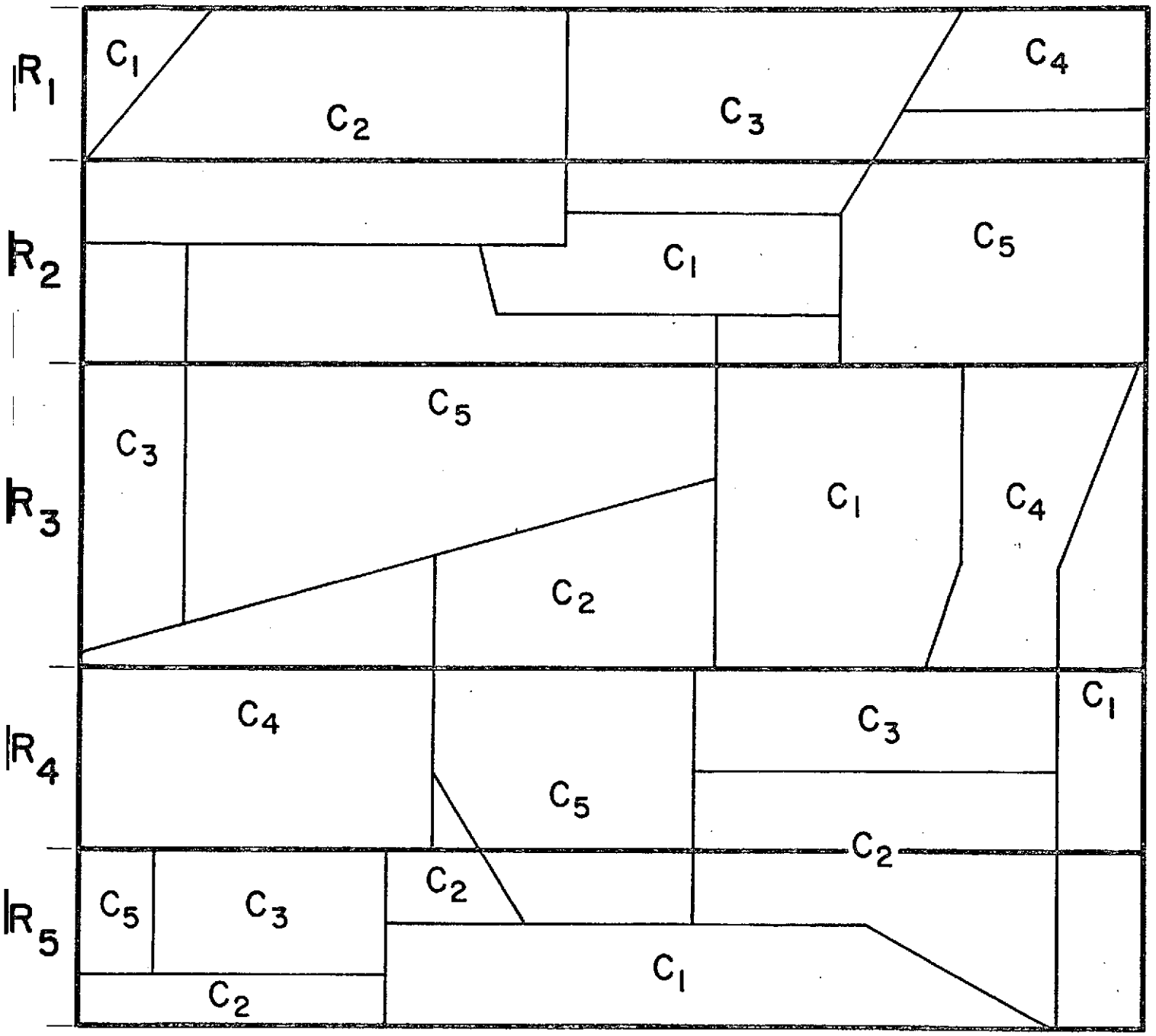


Figure 1: A subdivision of a region with five crops for a three-stage sampling plan.

ESTIMATION OF CROP ACREAGE THROUGH SAMPLING OF REMOTELY SENSED DATA: CLASSIFICATION ERROR MATRIX UNKNOWN*

R. S. Chhikara and P. L. Odell

The University of Texas at Dallas

*This research is carried out for NASA and supported by Contract NAS9-13512.

Estimation of Crop Acreage Through Sampling of Remotely Sensed Data

R. S. Chhikara and P. L. Odell
The University of Texas at Dallas

1. INTRODUCTION

In this report we consider the problem of estimating crop acreages in an area using samples from remotely sensed data for the area. Rationale for using samples is to avoid enormous cost and time that might be involved otherwise if the full data is processed. In particular, such would be the case for a ERTS scene which generally has over half a million data points and covers approximately an area of 100 X 100 square miles.

While considering crop acreage estimation, it is desirable to assume that the underlying region is an agricultural area and that every data point is identifiable with respect to certain known types of crops. In a larger context of an arbitrary area, it is sufficient for the condition of identification to hold with respect to the underlying known earth resources.

To formulate the problem, let $\Pi_1, \Pi_2, \ldots, \Pi_m$ denote the m crops in the area and $p_1, p_2, \ldots, p_m$ be the proportions of their acreages. Next, let $P(i/j)$ be the probability of classifying a resolution element (pixel) from Π_j into Π_i using a classification algorithm. Then associated with such classification algorithm processing of full remotely sensed data would amount to expecting proportion of Π_i acreage given by

$$e_i = \sum_{j=1}^{m} p_j \; P(i/j), \; i, j, 2, \ldots, m \qquad (1)$$

Equivalently,

$$e = Pp \qquad (2)$$

where

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_m \end{bmatrix}, \quad p = \begin{bmatrix} p_1 \\ p_2 \\ \cdot \\ \cdot \\ \cdot \\ p_m \end{bmatrix} \quad \text{and} \quad P = \begin{bmatrix} P(1/1) & P(1/2) \ldots & P(1/m) \\ P(2/1) & P(2/2) \ldots & P(2/m) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ P(m/1) & P(m/2) \ldots & P(m/m) \end{bmatrix}$$

If the vector e and matrix P are known, one gets p by solving (2) subject to $\Sigma p_i = 1$. Otherwise one needs to consider estimation of e or P or both of these as the case may be in order to ascertain about p.

In general, e will be unknown. An estimate of e is given by the vector of observed proportions of resolution elements processed and classified into Π_i, i = 1, 2, ..., m using sample data obtained under a sampling plan. Regarding P, two cases arise:

(i) P is known

(ii) P is unknown

Again, P will generally be unknown and for an estimate of P one would require some suitably selected amount of the ground truth, probably independent of the previous sample data used in estimating e. Moreover, care should be exercised in handling of the sampled ground truth. Most often, it would be desirable to use the sample for both training the classifier and obtaining estimate of P.

Clearly how much complex the estimation problem is depends upon how much one knows about P. In an entirely unknown situation of classification with untrained classifier, estimation of acreage proportions can be very misleading. As such we would assume for the classifier to be properly trained and thus are able to achieve reliable estimate of P.

2. PROPORTION ESTIMATES

Let $\hat{e}$ be an estimate of e and $\hat{P}$ of P. Then an estimate of p is given by

(i) $\hat{p} = P^{-1}\,\hat{e}$ when P is known (3)

and

(ii) $\hat{\hat{p}} = \hat{P}^{-1}\,\hat{e}$ when P is unknown (4)

In case (i) $\hat{p}$ is an unbiased estimate of p whenever e is an unbiased estimate of e. However, if estimate for both e and P are involved in estimating p as in case (ii), it may hardly be possible to obtain an unbiased estimate of p in $\hat{\hat{p}}$. Henceforth, we will assume $\hat{e}$, and so also $\hat{p}$, to be unbiased estimate and it is the vector of observed proportions of resolution elements processed and classified into different Π_i's (This will be the case due to the sampling plans being considered in this report.) Also, we would restrict our investigation to case (ii) and for case (i), refer to Chhikara and Odell [1]. Though final results in [1] pertain to specific sampling plans, the discussion there is of general nature and results can be adopted in any sampling situation.

3. MEAN SQUARE ERRORS OF ESTIMATES

First we calculate the bias of $\hat{\hat{p}}$ given by

$$\begin{aligned}\text{Bias } \{\hat{\hat{p}}\} &= E\,[\hat{\hat{p}}-p] \\ &= E\,[\hat{P}^{-1}\hat{e} - P^{-1}e] \\ &= E\,[\hat{P}^{-1}(\hat{e}-e) + (\hat{P}^{-1} - P^{-1})e] \\ &= E\,[\hat{P}^{-1} - P^{-1}]e \qquad (5)\end{aligned}$$

because the first term is zero due $E(\hat{e}-e) = 0$ for a given $\hat{P}^{-1}$. Clearly, the bias depends upon how much bias there is in $\hat{P}^{-1}$, and

$$\text{Bias } \{\hat{\hat{p}}\} = (\text{Bias } \{\hat{P}^{-1}\})e. \qquad (6)$$

In order to find the mean square error of any component of $\hat{\hat{p}}$, we first consider the evaluation of matrix,

$$\begin{aligned}E\,[(\hat{\hat{p}}-p)(\hat{\hat{p}}-p)^T] &= E\,[(\hat{P}^{-1}\hat{e} - P^{-1}e)(\hat{P}^{-1}\hat{e} - P^{-1}e)^T] \\ &= E\,[(\hat{P}^{-1})(\hat{e}-e)(\hat{e}-e)^T(\hat{P}^{-1})^T + (\hat{P}^{-1} - P^{-1})\, ee^T\, (\hat{P}^{-1} - P^{-1})^T] \\ &= E\,[(\hat{P}^{-1})\, M(\hat{P}^{-1})^T] + E\,[(\hat{P}^{-1} - P^{-1})\, ee^T\, (\hat{P}^{-1} - P^{-1})^T], \qquad (7)\end{aligned}$$

where M denotes the covariance matrix of $\hat{e}$.

Denoting the (i,j) th element of P^{-1} by P^{ij} and that of $\hat{P}^{-1}$ by $\hat{P}^{ij}$, it follows from (7) that the mean square error of $\hat{\hat{p}}_i$ is given by

$$E\left[\sum_{k=1}^{m} \hat{P}^{ik} \sum_{j=1}^{m} \hat{P}^{ij}\, \mathrm{Cov}(\hat{e}_j,\hat{e}_k) + \sum_{k=1}^{m} e_k(\hat{P}^{ik}-P^{ik}) \sum_{j=1}^{m} (\hat{P}^{ij}-P^{ij})\, e_j\right]$$

where E stands for expectation with respect to $\hat{P}$. Denoting it by MSE $\{\hat{\hat{p}}_i\}$, we have

$$\text{MSE}\,(\hat{\hat{p}}_i) = E\Bigg[\sum_{j=1}^{m} (\hat{P}^{ij})^2\,\text{Var}\,(\hat{e}_j) + \sum_{\substack{j=1\\ j\neq k}}^{m}\sum_{k=1}^{m} \hat{P}^{ij}\,\hat{P}^{ik}\,\text{Cov}\,(\hat{e}_j,\hat{e}_k) + \sum_{j=1}^{m} e_j^2\,E\,[\hat{P}^{ij} - P^{ij}]^2 + \sum_{\substack{j=1\\ j\neq k}}^{m}\sum_{k=1}^{m} e_j e_k E[(\hat{P}^{ij}-P^{ij})(\hat{P}^{ik}-P^{ik})]\Bigg], \tag{8}$$

$i = 1, 2, \ldots, m.$

The procedure for obtaining $\hat{P}$ would involve sampling of ground truth independent of samples taken for estimating e. In the following section we give expressions for $v(\hat{e}_i)$ and Cov $(\hat{e}_i, \hat{e}_j)$ considering different sampling plans. To evaluate expectation in (8), one needs to find the distribution of $\hat{P}$. This will, of course, depend upon how $\hat{P}$ is obtained. In general, it will be difficult to obtain any exact distribution of $\hat{P}$. However, if the sampling of ground truth involves separate independent samples from each crop and $\hat{P}$ is obtained as the matrix of observed proportions among randomly selected pixels classified into different crops using a classifier, each column vector of $\hat{P}$ has a multinomial distribution and is stochaotically independent of the others in $\hat{P}$. Since expectation in (8) is for elements of $\hat{P}^{-1}$, it may not be easy to derive the MSE $\{\hat{\hat{p}}_i\}$ in a closed form, especially if the number of classes is large. As such we now consider the two-class case and show the procedure for obtaining Bias $\{\hat{\hat{p}}_i\}$ and MSE $\{\hat{\hat{p}}_i\}$, i=1, 2.

Two-class Problem

Often interest lies in ascertaining acreage of a specified crop in an area. In view of this one may consider Π_1 to be the crop of main interest and Π_2 to be its complementary part. Without loss of generality, let us assume density functions for Π_1 and Π_2 to be symmetric about certain

location points. Then P(1/2) is same as P(2/1). Denoting this common value by Φ, one gets

$$P = \begin{bmatrix} 1-\Phi & \Phi \\ \Phi & 1-\Phi \end{bmatrix}$$

It now follows from (2) that

$$e_1 = p_1 + (p_2 - p_1)\,\Phi$$

and

$$e_2 = p_2 + (p_1 - p_2)\,\Phi \;, \quad (e_1 + e_2 = 1)$$

Assuming $\Phi \neq 1/2$, we have

$$p_1 = \frac{e_1 - \Phi}{1 - 2\,\Phi}$$

and

$$p_2 = \frac{e_2 - \Phi}{1 - 2\,\Phi}$$

Considering $\hat{e}_1$, $\hat{e}_2$ and $\hat{\Phi}$ ($\hat{\Phi} \neq 1/2$) estimates for e_1, e_2 and Φ, respectively, obtained according to the procedure outlined in the previous paragraphs, one has

$$\hat{p}_1 = \frac{\hat{e}_1 - \hat{\Phi}}{1 - 2\,\hat{\Phi}}$$

$$\hat{p}_2 = \frac{\hat{e}_2 - \hat{\Phi}}{1 - 2\,\hat{\Phi}}$$

Clearly, $\hat{p}_1 + \hat{p}_2 = 1$ because of $\hat{e}_1 + \hat{e}_2 = 1$. Next, it follows that for $k = 1, 2$,

$$\text{Bias}\,\{\hat{p}_k\} = (e_k - 1/2)\,(E[T] - \Theta)$$

and

$$\text{MSE}\,\{\hat{p}_k\} = \text{Var}(\hat{e}_k)\,E[T^2] + (e_k - 1/2)^2\,E[T - \Theta]^2$$

where

$$T = (1 - 2\,\hat{\Phi})^{-1} \qquad \text{and} \qquad \Theta = (1 - 2\,\Phi)^{-1}$$

To obtain $E[T]$ and $E[T^2]$, one way is to find the distribution of $\hat{\Phi}$. Let r_1 out of N_1 pixels sampled from Π_1 be misclassified into Π_2 and r_2 out of N_2 pixels sampled from Π_2 be misclassified into Π_1. Then

$$\hat{\Phi} = \frac{r_1 + r_2}{N_1 + N_2}$$

provides an estimate of Φ. The condition of $\hat{\Phi} \neq 1/2$ is easily met if we restrict $N_1 + N_2$ to an odd number, a mild restriction which should not undermine the generality of present discussion. Denoting $r = r_1 + r_2$ and $N = N_1 + N_2$, the random variable r has a Binomial distribution with proportion Φ and sample size N. Accordingly

$$E[T] = \sum_{r=0}^{N} \frac{N}{N - 2r} \binom{N}{r} \Phi^r (1 - \Phi)^{N-r}$$

and

$$E[T^2] = \sum_{r=0}^{N} \frac{N^2}{(N - 2r)^2} \binom{N}{r} \Phi^r (1 - \Phi)^{N-r}$$

However, it is difficult to give any closed-form expression for $E[T]$ and $E[T^2]$. Due to $-1 < (1 - \frac{2r}{N}) < 1$,

$$E[T] = \sum_{s=0}^{\infty} (2/N)^s \mu_s$$

and

$$E[T^2] = \sum_{s=0}^{\infty} (s + 1)(2/N)^s \mu_s$$

where μ_s is the sth moment of Binomial distribution about the origin and can be easily obtained by evaluating the sth derivative of the moment generating function at the origin, i.e.

$$\mu_s = \left[\frac{\partial^s}{\partial t^s} (1 - \Phi + \Phi e^t)^N\right]_{t=0} .$$

It can be shown that

$$E[T] = (1 - 2\Phi)^{-1} + O(1/N)$$

$$E[T^2] = (1 - 2\Phi)^{-2} + O(1/N) .$$

Then asymptotically, i.e. as N becomes large,

$$\text{Bias } \{\hat{p}_k\} = 0$$

and

$$\text{MSE } \{\hat{p}_k\} = \text{Var } (\hat{p}_k) = (1 - 2\Phi)^{-2} \text{ Var}(\hat{e}_k) .$$

4. SAMPLING PLAN AND COVARIANCE MATRIX OF $\hat{e}$

In a remote sensing situation involving collection of data over a large region, we suggest a stratified random sampling scheme with three stages. First of all, stratification will be most effective if strata are formed on the basis of at least

(i) predominance of various crops,

(ii) latitude,

(iii) longitude.

The first factor would lead to homogeneity in various strata whereas the other two factors are important from the point of assessing P appropriately.

Let R denote the region for which crop acreages are to be estimated. With the consideration of (i) - (iii), suppose R is stratified into strata

R_{st}, s=1,2,...,a and t=1,2,...,b, with weights w_{st}, the proportion of acreage (number of pixels in a stratum divided by the total number of pixels in the region), i.e.

$$R = \bigcup_{s,t} R_{st}$$

with

$$1 = \sum_{s,t} w_{st} .$$

Due to (i), one may expect elimination of many scenes or even many strata if interest lies in only a few specific types of crops. Nevertheless, for sampling purposes, select ERTS scenes at the first stage, strips within scenes at the second stage and scanlines within strips at the third stage. Of course, one can go one more stage in selecting pixels within scanline. However, this would not be convenient as far as processing is concerned. As such, this stage is not considered for the sampling.

For stratum R_{st}, we denote the following:

$e_{st\ ijhk}$ = expected proportions of pixels in π_i for kth scanline in hth strip of jth scene,

$e_{st\ ijh}$ = expected proportions of pixels in π_i for hth strip in jth scene,

$e_{st\ ij}$ = expected proportion of pixels in π_i for jth scene,

$e_{st\ i}$ = expected proportion of pixels in π_i.

Then

$$e_i = \sum_{s=1}^{a}\sum_{t=1}^{b} w_{st}\, e_{st\ i}, \qquad i = 1,2,...,m.$$

Next, let G_{st}, H_{st}, R_{st} and n denote the number of scenes, number of strips per scene, number of scanlines per strip and number of pixels per

scanline, respectively, for stratum R_{st}. Suppose g_{st}, h_{st} and r_{st} are the corresponding number of scenes, number of strips in a scene, number of scanlines in a strip that are selected randomly at three stages. Let $n_{st\ ijhk}$ be the number of pixels classified into π_i for kth selected scanline in hth selected strip of jth selected scene. Then considering the observed proportions for estimates, one has

$$\hat{e}_{st\ ijhk} = \frac{n_{st\ ijhk}}{n}$$

$$\hat{e}_{st\ ijh} = \frac{1}{nr_{st}} \sum_{k=1}^{r_{st}} n_{st\ ijhk}$$

$$\hat{e}_{st\ ij} = \frac{1}{nr_{st}h_{st}} \sum_{h=1}^{h_{st}} \sum_{k=1}^{r_{st}} n_{st\ ijhk}$$

$$\hat{e}_{st\ i} = \frac{1}{nr_{st}h_{st}g_{st}} \sum_{j=1}^{g_{st}} \sum_{h=1}^{h_{st}} \sum_{k=1}^{r_{st}} n_{st\ ijhk}$$

and

$$\hat{e}_i = \sum_{s=1}^{a} \sum_{t=1}^{b} w_{st}\ \hat{e}_{st\ i},\ i=1,2,\ldots,m.$$

For the covariance matrix,

$$\mathrm{Var}(\hat{e}_{st\ i}) = \left(1 - \frac{g_{st}}{G_{st}}\right) \frac{1}{g_{st}(G_{st}-1)} \sum_{j=1}^{G_{st}} (e_{st\ ij} - e_{st\ i})^2$$

$$+ \left(1 - \frac{h_{st}}{H_{st}}\right) \frac{1}{g_{st}h_{st}G_{st}(H_{st}-1)} \sum_{j=1}^{G_{st}} \sum_{h=1}^{H_{st}} (e_{st\ ijh} - e_{st\ ij})^2$$

$$+ \left(1 - \frac{r_{st}}{R_{st}}\right) \frac{1}{g_{st}h_{st}r_{st}G_{st}H_{st}(R_{st}-1)} \sum_{j=1}^{G_{st}} \sum_{h=1}^{H_{st}} \sum_{k=1}^{R_{st}} \left(e_{st\ ijhk} - e_{st\ ijh}\right)^2$$

$$\mathrm{Cov}(\hat{e}_{st\ i}, \hat{e}_{st\ i'}) = \left(1 - \frac{g_{st}}{G_{st}}\right) \frac{1}{g_{st}(G_{st}-1)} \sum_{j=1}^{G_{st}} (e_{st\ ij} - e_{st\ i})(e_{st\ ij} - e_{st\ i'})$$

$$+ \left(1 - \frac{h_{st}}{H_{st}}\right) \frac{1}{g_{st}h_{st}G_{st}(H_{st}-1)} \sum_{j=1}^{G_{st}} \sum_{h=1}^{H_{st}} (e_{st\ ijh} - e_{st\ ij})(e_{st\ ijh} - e_{st\ i'j})$$

$$+ \left(1 - \frac{r_{st}}{R_{st}}\right) \frac{1}{g_{st}h_{st}r_{st}G_{st}H_{st}(R_{st}-1)} \sum_{j=1}^{G_{st}} \sum_{h=1}^{H_{st}} \sum_{k=1}^{R_{st}} (e_{stijhk} - e_{stijh})(e_{stijhk} - e_{sti'jh}).$$

Thus the covariance matrix for $\hat{e}$ can be obtained because

$$\mathrm{Var}(\hat{e}_i) = \sum_{s=1}^{a} \sum_{t=1}^{b} w_{st}^2 \, \mathrm{Var}(\hat{e}_{st\ i}) \tag{9}$$

$$\mathrm{Cov}\ (\hat{e}_i, \hat{e}_{i'}) = \sum_{s-1}^{a} \sum_{t=1}^{b} w_{st}^2 \ \mathrm{Cov}(\hat{e}_{st\ i}, \hat{e}_{st\ i'}) \ . \tag{10}$$

For an estimate of the covariance matrix,

$$\hat{\mathrm{Var}}(\hat{e}_i) = \sum_{s=1}^{a} \sum_{t=1}^{b} w_{st}^2 \ \hat{\mathrm{Var}}(\hat{e}_{st\ i})$$

$$\hat{\mathrm{Cov}}\ (\hat{e}_i, \hat{e}_{i'}) = \sum_{s=1}^{a} \sum_{t=1}^{b} w_{st}^2 \ \hat{\mathrm{Cov}}(e_{st\ i}, \hat{e}_{st\ i'})$$

where

where

$$\hat{\text{Var}}(\hat{e}_{st\ i}) = (1-\frac{g_{st}}{G_{st}})\frac{1}{g_{st}(g_{st}-1)}\sum_{j=1}^{g_{st}}(\hat{e}_{st\ ij}-\hat{e}_{st\ i})^2$$

$$+ (1-\frac{h_{st}}{H_{st}})\frac{1}{G_{st}g_{st}h_{st}(h_{st}-1)}\sum_{j=1}^{g_{st}}\sum_{h=1}^{h_{st}}(\hat{e}_{st\ ijh}-\hat{e}_{st\ ij})^2$$

$$+ (1-\frac{r_{st}}{R_{st}})\frac{1}{G_{st}H_{st}g_{st}h_{st}r_{st}(h_{st}-1)}\sum_{j=1}^{g_{st}}\sum_{h=1}^{h_{st}}\sum_{k=1}^{r_{st}}(\hat{e}_{st\ ijhk}-\hat{e}_{st\ ijh})^2. \quad (11)$$

Similarly one can write $\hat{\text{Cov}}(\hat{e}_{st\ i}, \hat{e}_{st\ i'})$ by replacing the sum of squares terms in the variance by the corresponding sum of product terms.

5. OPTIMUM SAMPLE SIZE

Taking the cost factor into consideration one may want to know the sample size that either minimizes the cost for a specified mean square error or minimize the mean square error for a given cost. In remote sensing, the sampling cost would mainly involve a large initial cost plus the processing cost. Although it would depend upon a situation, the cost function may be considered as

$$C_{st} = C_1 g_{st} + C_2 g_{st} h_{st} + C_3 g_{st} h_{st} r_{st}$$

for when sampling in stratum R_{st}, and

$$C = C_1 \Sigma g_{st} + C_2 \Sigma g_{st} h_{st} + C_3 \Sigma g_{st} h_{st} r_{st}$$

for the whole region.

In the present context, there is an additional cost of taking samples for estimating P and this can be expressed in the form of $C' = \sum_{i=1}^{m} C_{0i} n_i$. Thus the overall cost involved is given by

$$C'' = C' + C .$$

Next, the mean square errors, $MSE\{\hat{\hat{p}}_k\}$, k=1,2,...,m, are obtained from (8) after making substitution from (9) and (10). Now, if the cost is fixed, say $C'' \leq C_0$, a determination of sample sizes, n_i's, g_{st}'s, g_{st}'s, h_{st}'s and r_{st}'s, can be achieved by solving equations obtained by equating the partial derivatives of $MSE\{\hat{\hat{p}}_k\} + \lambda(C''-C_0)$, k=1,2,...,m and λ a Lagrange multiplier, with respect to n_i's, g_{st}'s, h_{st}'s and r_{st}'s to zero. Similarly, for any fixed values, say σ_k^2, for $MSE\{\hat{\hat{p}}_k\}$, k=1,2,...,m, this can again be achieved by considering the function $C'' + \lambda_K(MSE\{\hat{\hat{p}}_k\} - \sigma_k^2)$, k=1,2,...,m, for minimization. Of course, this procedure may lead to k different values for various sample sizes. For a unique determination, take the largest value in each case.

It may be noted that under this procedure, it is not possible to give any closed form expression for any sample size and its carrying out would involve some optimization technique.

One direct way to simplify the problem is to treat the two types of cost separately. For example, if we consider only the cost C and variances given in (9) for the purpose of determining g_{st}'s, h_{st}'s and r_{st}'s, the problem is greatly simplified. Denoting

$$(G_{st}-1)\, S^2_{st\ 1} = \sum_{j=1}^{G_{st}} (e_{st\ ij} - e_{st\ i})^2$$

$$G_{st}(H_{st}-1)S^2_{st\ 2} = \sum_{j=1}^{G_{st}} \sum_{h=1}^{H_{st}} (e_{st\ ijh}-e_{st\ ij})^2$$

and

$$G_{st}H_{st}(R_{st}-1)S^2_{st\ 3} = \sum_{k=1}^{R_{st}} \sum_{h=1}^{H_{st}} \sum_{j=1}^{G_{st}} (e_{st\ ijhk}-e_{st\ ijh})^2 .$$

(9) may be written as

$$\mathrm{Var}(\hat{e}_i) = \sum_{s,t} w^2_{st}\left[\frac{1}{g_{st}}(S^2_{st\ 1}-S^2_{st\ 2}/H_{st}) + \frac{1}{g_{st}h_{st}}(S^2_{st\ 2}-S^2_{st\ 3}/R_{st}) + S^2_{st\ 3}/g_{st}h_{st}r_{st} - S^2_{st\ 1}/G_{st}\right].$$

which is a function of g_{st}, $g_{st}h_{st}$ and $g_{st}h_{st}r_{st}$ as is the case with the above cost function. Thus, the quantity $\mathrm{Var}(\hat{e}_i) + \lambda(C-C_0)$ or $C + \lambda_i(\mathrm{Var}(\hat{e}_i) - V_i^2)$, and equivalently $\mathrm{Var}(\hat{e}_i)$ for fixed C or C for fixed Var $(\hat{e}_i)$, is minimized when

$$g_{st} = w_{st}\sqrt{(S^2_{st\ 1}- S^2_{st\ 2}/H_{st})/\lambda C_1}$$

$$g_{st}h_{st} = w_{st}\sqrt{(S^2_{st\ 2} - S^2_{st\ 3}/R_{st})/\lambda C_2}$$

$$g_{st}j_{st}l_{st} = w_{st}\sqrt{S^2_{st\ 3}/\lambda C_3} .$$

Accordingly,

$$r_{st} = S_{st\ 3}\sqrt{C_2}/\sqrt{C_3(S^2_{st\ 2}-S^2_{st\ 3}/R_{st})}$$

$$h_{st} = \sqrt{C_1(S^2_{st\ 2}-S^2_{st\ 3}/R_{st})/C_2(S^2_{st\ 1}-S^2_{st\ 2}/H_{st})}$$

and $$g_{st} \propto w_{st} \sqrt{(S^2_{st\,1} - S^2_{st2}/H_{st})/C_1} \,.$$

These values are, of course, for the case of $\mathrm{Var}(\hat{e}_i)$. Similarly, one can obtain sample sizes corresponding to other variances. By choosing the maximum of these values in different cases, a unique solution can be obtained.

6. FURTHER COMMENTS

When stratification is based upon factors of latitude and longitude, one might expect different P's over different strata. However, if proper adjustment can be made in the classification algorithm with respect to spectral variation so that P remains the same, there is no need to make any change in the above procedure. On the other hand, one should find both the actual proportion estimate and its mean square error separately for each stratum and then by combining these one is able to obtain so called $\hat{\hat{p}}$ and its mean square error. In this situation the formula in (8) is still valid but stratumwise.

Our discussion given in Section 1-3 is quite general and can be specialized and different sampling schemes can be adopted. For example, if our inference set is only one scene, then proportion estimates and their mean square errors are again obtained as discussed in Section 1-3, but as to sampling scheme there may not be any need of stratification and the sampling is done in two stages rather than in three stages.

REFERENCES

[1] Chhikara, R.S. and Odell, P.L. "Acreage Estimates for Crops Using Remote Sensing Techniques," Preliminary Report submitted to NASA, 1973.

[2] Cochran, W. Sampling Techniques, 2nd Edition, John Wiley and Sons, Inc., New York, 1963.

ON COMBINING POPULATIONS IN STATISTICAL CLASSIFICATION USING REMOTE SENSING DATA

J. P. Basu

and

P. L. Odell

The University of Texas at Dallas

This research is supported by Johnson Space Research Center Contract #NAS9-13512.

ABSTRACT

Observations may be known to be coming from m+1 normal populations having same dispersion matrix; but one may be interested in identifying observations from π_0 only. Then the usual practice is to merge the other populations $\pi_1,\ldots,\pi_m$ into one single population π and assume it to be normal. This is done in order to achieve the computational convenience of two population classification. This paper investigates the effectiveness of such practice. It has been found that in some situations this practicle may decrease the proper classification probability of observations from π_0 considerably.

Key Words

Dispersion matrix, Bayes' procedure, misclassification probability, proper classification probability, prior probability, misclassification cost.

1. Introduction

Let us suppose that all of our observations come from (m+1) p-variate normal populations $\pi_0,\pi_1,\pi_2,\ldots,\pi_m$ with densities $N_p(\mu_i,V)(i=0,1,2,\ldots,m)$, where $\mu_i^T = (\mu_{i1},\mu_{i2},\ldots,\mu_{ip})$ is a vector of p means of the populations π_i and V is the dispersion matrix, same for all populations. In general the problem of classification into more than two classes (m+1>p>1) is more complex and computationally inconvenient than the problem of classification into two classes. In some situations, instead of finding which population an observation has come from, one may be interested in finding only if the observation has come from some particular population, say π_0. There, for achieving the convenience and simplicity of two-population classification problem, one may think of merging the populations $\pi_1,\pi_2,\ldots,\pi_m$ into one single population π and further assume the resulting single population π to be normal. Then the question that needs to be answered is how badly the classification procedure based on such assumptions affects the misclassification probabilities. This paper endeavors to answer this question.

In Bayes' procedure the prior probabilities play a significant role. Improper choice of prior probabilities affects the misclassification probabilities. The usual practice in remote sensing data analysis is to assume the prior probabilities q_0 and q of μ_0 and μ to be equal, even though it has been assumed that the prior probabilities $q_0,q_1,\ldots,q_m$ of $\pi_0,\pi_1,\ldots,\pi_m$ are equal. Therefore in answering the question raised in the above paragraph we should take into consideration our assumption about the prior probabilities.

Thus our study will be concerned with combination of following sets of assumptions.

Assumptions regarding population densities

(1) Before merger the populations $\pi_0, \pi_1, \ldots, \pi_m$ have densities $N_p(\mu_i, V)$, $i=0,1,\ldots,m$;

(2) After merger, π_0 has density $N_p(\mu_0, V)$ and the density of π is a linear combination of the densities $N_p(\mu_i, V)$, $i=1,\ldots,m$;

(3) After merger π_0 has density $N_p(\mu_0, V)$ and π has a density $N_p(\mu, V')$, where μ and V' will be specified later.

Assumptions regarding prior probabilities

(a) $q_0 = 1/(m+1)$, $q_1 = q_2 = \ldots = q_m = 1/(m+1)$;

(b) $q_0 = 1/2$, $q_1 = q_2 = \ldots = q_m = 1/2m$.

2. Dispersion Matrix of the Population π

The true density $p(x)$ of the population π is given by

$$p(x) = \sum_{i=1}^{m} [q_i/(1-q_0)]\, N_p(\mu_i, V). \tag{1}$$

Now since in both assumptions (a) and (b) we have

$$q_1 = q_2 = \ldots = q_m$$

and

$$q_i/(1-q_0) = q_1/(1-q_0) = 1/m,$$

then the density $p(x)$ under (a) or (b) is given by

$$p(x) = (1/m) \sum_{i=1}^{m} N_p(\mu_i, V) . \tag{2}$$

Therefore the mean μ of π is given by

$$\mu = E_\pi(X) = \int_{R_p} x\, p(x)\, dx = \frac{1}{m} \sum_{i=1}^{m} \int_{R_p} x\, N_p(\mu_i, V)\, dx$$

$$= \frac{1}{m} (\mu_1 + \ldots + \mu_m) = \bar{\mu} , \tag{3}$$

and the dispersion matrix V' of π is given by

$$V' = E_{\pi}\,[(X-\bar{\mu})(X-\bar{\mu})^T]$$

$$= \frac{1}{m}\sum_{i=1}^{m} \int_{R_p} (x-\bar{\mu})(x-\bar{\mu})^T N_p(\mu_i, V)\, dx$$

$$= \frac{1}{m}\sum_{i=1}^{m} \int_{R_p} \{(x-\mu_i)(x-\mu_i)^T + 2(\mu_i-\bar{\mu})^T(x-\mu_i) + (\mu_i-\bar{\mu})(\mu_i-\bar{\mu})^T\} N_p(\mu_i, V)\, dx$$

$$= V + \frac{1}{m}\sum_{i=1}^{m} (\mu_i-\bar{\mu})(\mu_i-\bar{\mu})^T . \qquad (4)$$

Here $E_{\pi}(\cdot)$ denotes expectation of the population π.

Therefore, if we want to use a normal density function for π instead of the density $p(x)$ given by (2), we have to use $N_p(\bar{\mu}, V')$, where V' is given by (4).

Expression for $(V')^{-1}$

Let us write $V_0 = V$ and

$$V_i = V_{i-1} + \frac{1}{m}(\mu_i-\bar{\mu})(\mu_i-\bar{\mu})^T, \quad (i=1,2,\ldots,m). \qquad (5)$$

Then $V_m = V'$. We know (Rao, 1965; p. 29) that if A is a nonsingular $p \times p$ matrix and α and β are two $p \times 1$ vectors, then

$$(A + \alpha\beta^T)^{-1} = A^{-1} - \frac{A^{-1}\alpha\beta^T A^{-1}}{1+\beta^T A^{-1}\alpha} . \qquad (6)$$

Thus, $$V_1^{-1} = V^{-1} - \frac{1}{m}V^{-1}(\mu_1-\bar{\mu})(\mu_1-\bar{\mu})^T V^{-1}/\{1+(\mu_1-\bar{\mu})^T V^{-1}(\mu_1-\mu_2)\}, \qquad (7)$$

and $$V_i^{-1} = V_{i-1}^{-1} - \frac{1}{m}V_{i-1}^{-1}(\mu_i-\bar{\mu})(\mu_i-\bar{\mu})^T V_{i-1}^{-1}/\{1+(\mu_i-\bar{\mu})^T V_{i-1}^{-1}(\mu_i-\bar{\mu})\} ,$$

$(i=2,\ldots,m)$. Therefore

$$(V')^{-1} = V^{-1} - M \text{ (say)}. \qquad (8)$$

The exact expression for M that can be found recursively from (7) is not a simple one.

3. Acceptance Region For π_0: Assumptions (1) and (2)

We will denote the acceptance region of π_0 given by a Bayes' classification procedure by B. B_{2a}, for example, will denote the region B obtained under the assumptions (2) and (a). If the populations $\pi_0,\pi_1,\ldots,\pi_m$ have densities $p_0(x)$, $p_1(x),\ldots,p_m(x)$ and prior probabilities $q_0,q_1,\ldots,q_m$ and $C(\pi_j|\pi_i),(j\neq i)$, denote the cost of misclassifying an observation from π_i into π_j, then it is well known (Anderson, 1958; p. 143) that the acceptance region B for π_0 is given according to Bayes' procedure by

$$B=\cap_{j=1}^{m}\{X:\sum_{i=1}^{m}q_ip_i(X)C(\pi_0|\pi_i)<\sum_{i\neq j}q_ip_i(X)C(\pi_j|\pi_i)+q_0p_0(X)C(\pi_j|\pi_0)\}. \tag{9}$$

In case of assumption (1) if we assume all costs of misclassification to be equal, then we have

$$\begin{aligned} B_1 &= \cap_{j=1}^{m}\{x:\ q_0p_0(x) > q_jp_j(x)\} \\ &= \{x:\ q_0p_0(x) > \max_{1\leq j\leq m}[q_jp_j(x)]\}\ . \end{aligned} \tag{10}$$

Therefore,
$$B_{1a} = \{x:\ p_0(x) > \max_{1\leq j\leq m} p_j(x)\} \tag{11}$$

and
$$B_{1b} = \{x:\ p_0(x) > (1/m)\max_{1\leq j\leq m} p_j(x)\}\ . \tag{12}$$

In case of assumption (2) we assume that

$$C(\pi_0|\pi_i) = C(\pi_i|\pi_0) = C(\pi_j|\pi_0), \quad \text{for all } i\neq j\neq 0$$

and
$$C(\pi_i|\pi_j) = 0, \quad \text{for all } i,j\neq 0\ .$$

Then we have

$$\begin{aligned} B_2 &= \{x:\ q_0p_0(x) > q_1p_1(x) + \ldots + q_mp_m(x)\} \\ &= \{x:\ q_0p_0(x) > (1-q_0)\sum_{i=1}^{m}[q_i/(1-q_0)]p_i(X)\}\ . \end{aligned} \tag{13}$$

Therefore,
$$B_{2a} = \{x:\ p_0(x) > p_1(x) + \ldots + p_m(x)\} \tag{14}$$

and
$$B_{2b} = \{x:\ p_0(x) > (1/m)\ [p_1(x) + \ldots + p_m(x)]\}\ . \tag{15}$$

Now since $\max\{p_1(X),\ldots,p_m(x)\} \geq (1/m)[p_1(x)+\ldots+p_m(x)]$, the following set theoretic inequalities are evident.

$$B_{1b} \supset B_{2b} \supset B_{1a} \supset B_{2a} \ . \tag{16}$$

When $p_i(x) = N_p(\mu_i, V)$, $(i=0,1,\ldots,m)$, then we know (Anderson, 1958; 147) that B_i can be expressed as follows.

$$B_1 = \bigcap_{j=1}^{m} \{x: W_j(x) \geq \log(q_j\ q_0)\}$$

$$= \{x: W_j(x) > \log(q_j|q_0),\ \text{for } j=1,\ldots,m\} \ ,$$

$$\text{where } W_j(x) = (\mu_0-\mu_j)^T V^{-1}[x - (1/2)(\mu_0 + \mu_j)] \ . \tag{18}$$

$$\text{Therefore} \qquad B_{1a} = \{x: W_j(x) \geq 0,\ \text{for } j=1,\ldots,m\} \tag{19}$$

$$\text{and} \qquad B_{1b} = \{x: W_j(x) \geq -\log m,\ \text{for } j=1,\ldots,m\} \ .$$

When the means $\mu_1,\mu_2,\ldots,\mu_m$ are collinear, without loss of generality we can assume the populations to be two dimensional and assume the means to be as follows:

$$\mu_0 = \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} \ , \quad \mu_i = \begin{bmatrix} a_i \\ 0 \end{bmatrix} \ , \quad (i = 1,2,\ldots,m). \tag{20}$$

In case of three populations, we can take

$$\mu_0 = \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} \ , \quad \mu_1 = \begin{bmatrix} -a \\ 0 \end{bmatrix} \ \text{and} \ \ \mu_2 = \begin{bmatrix} a \\ 0 \end{bmatrix} \ . \tag{21}$$

Let the 2 x 2 disperion matrix V be given by $V = [v_{ij}]$. Then we have

$$W_1 = (1/|V|)\ [x_1-(1/2)(a_0-a)][(a_0+a)\ v_{22} - b_0 v_{12}]$$

$$+ (1/|V|)[x_2 - (1/2)b_0][b_o v_{11} - v_{12}(a_0+a)] \tag{22}$$

and $$W_2 = (1/|V|)\ [x_1 - (1/2)(a_0+a)][(a_0-a)\ v_{22} - b_0\ v_{12}]$$
$$+ (1/|V|)\ [x_2 - (1/2)b_0][b_0\ v_{11} - v_{12}(a_0-a)]\ . \qquad (23)$$

The acceptance region B_1 for π_0 will be bounded by two straight lines, viz, $W_1=d$ and $W_2=d$, where d=0 in case (a) and d= -log 2 in case (b). The point of intersection of these two lines will be given by

$$(v_{12}C/D,\quad v_{22}C/D), \qquad (24)$$

where $$C = [\{(a_0^2 - a^2)\ v_{22} - 2a_0b_0v_{12} + b_0^2v_{11}\} + 2d|V|]$$

and $$D = 2b_0(v_{11}v_{22} - v_{12}^2) = 2b_0|V|,$$

$|V|$ denoting the determinant of V.

From (14) and (15) it follows that the acceptance region B_{2a} and B_{2b} can be given the following general description.

$$B_2 = \{x:\ p_o(x) > (1/\alpha)\ [p_1(x) + .. + p_m(x)]\},$$

where α should be taken as 1 for B_{2a} and as m for B_{2b}. In case of three populations with means given by (21) we may note that the condition

$$\alpha p_o(x) > p_1(x) + p_2(x)$$

is equivalent to

$$\log\ [\alpha\ p_o(x)/p_1(x)] > \log\ [p_2(x)/p_1(x) + 1].$$

Noting that $\mu_1 = -\mu_2$, we can write the above condition as

$$\log \alpha + (\mu_2+\mu_o)^T\ V^{-1}\ [x+(\mu_2-\mu_o)/2] > \log\ [1 + \exp\ (2\mu_2^TV^{-1}x)] \qquad (25)$$

Here α should be taken as 1 for B_{2a} and as 2 for B_{2b}.

When all means are collinear, without loss of generality we can assume the populations to be univariate and $\mu_1<\mu_2<\ldots<\mu_m$. V in this case is the common variance of the univariate populations. Now from (18) we have

$$W_j(x) = [(\mu_0-\mu_j)/V]\ [x-(1/2)\ (\mu_0+\mu_j)].$$

It now follows from (19) that if $\mu_0 < \mu_1$, then

$$B_{1a} = \{x: x < (\mu_0+\mu_1)/2\} , \tag{26}$$

if $\mu_i < \mu_0 < \mu_{i+1}$, then

$$B_{1a} = x:\{(\mu_0+\mu_i)/2 < x < (\mu_0+\mu_{i+1})/2\} \tag{27}$$

and if $\mu_0 > \mu_m$, then

$$B_{1a} = \{x: x > (\mu_0+\mu_m)/2\} . \tag{28}$$

The positions of the end points of the interval B_{2a} depend on Vlogm and the values of $\mu_0,\mu_1,\ldots,\mu_m$. B_{2a} can not be given a general description like B_{1a}.

4. Acceptance Region for π_0: Assumption (3)

When m populations $\pi_1,\ldots,\pi_m$ have been merged to form the population π, according to assumption (a) the prior probabilities q_0 and q of π_0 and π are then given by $q_0 = 1/(m+1)$ and $q = m/(m+1)$ and according to assumption (b) they are given by $q_0 = 1/2$ and $q = 1/2$. When π is assumed to have density $N_p(\bar{\mu},V')$, the acceptance region B_{3a} of π_0 will be given according to Bayes' procedure by

$$\begin{aligned} B_{3a} &= \{x: N_p(\mu_0,V) > mN_p(\bar{\mu},V')\} \\ &= \{x: -(x-\bar{\mu})^T(V')^{-1}(x-\bar{\mu}) + (x-\mu_0)^T V^{-1}(x-\mu_0) \\ &\qquad < \log(|V'|/|V|) - 2\log m\}. \end{aligned}$$

Using (8) we can write

$$B_{3a} = \{x: Q(x) < \log(|V'|/|V|) - 2\ \log m\} , \tag{29}$$

where $$Q(x) = (x-\bar{\mu})^T M(x-\bar{\mu}) - 2(\mu_0-\bar{\mu})^T V^{-1}\ [x-(\mu_0+\bar{\mu})/2]. \tag{30}$$

We can similarly have

$$B_{3b} = \{x: Q(x) < \log(|V'|/|V|)\} . \tag{31}$$

Obviously $B_{3b} \supset B_{3a}$.

Special Case

Let us suppose that the populations are two dimensional, their means are given by (20) subject to the condition $\bar{\mu} = 0$ and their common dispersion matrix is given by $V = \text{diag}\{v_1, v_2\}$. Then from (4) we have

$$V' = V + \sum_{i=1}^{m} \mu_i \mu_i^T/m = \text{diag}\{v_1 + \sum \mu_i \mu_i^T/m,\ v_2\}. \tag{32}$$

Therefore (33)

$$|V'|/|V| = 1 + \sum \mu_i^T \mu_i/mv_1$$

Writing

$$\omega^T = [(\Sigma a_i^2/m)^{\frac{1}{2}}\ 0],$$

we can express V' as

$$V' = V + \omega\omega^T.$$

From (6) and (8) we now have

$$M = (m_{ij}) = V^{-1}\omega\omega^T V^{-1}/\{1+\omega^T V^{-1}\omega\}.$$

Obviously M is a matrix such that

$$m_{11} = \Sigma a_i^2/\{v_1(mv_1 + a_i^2)\}$$

and $$m_{12} = m_{21} = m_{22} = 0.$$

Therefore, from (20) and (30) we have

$$Q(X) = m_{11}x_1^2 - 2\mu_o^T V^{-1}x + \mu_o^T V^{-1}\mu_o$$
$$= m_{11}x_1^2 - 2a_o x_1/v_1 - 2b_o x_2/v_2 + (a_o^2/v_1 + b_o^2/v_2)$$

From (29) and (31) it now follows that the regions B_{3s} $(s = a, b)$ are bounded by

$$\{x_1 - (a_o/v_1 m_{11})\}^2 = 2(b_o/m_{11}v_2)\{x_2 - A + (v_2/b_o)\log\alpha\} \tag{34}$$

where α should be taken as m for B_{3a} and as 1 for B_{3b} and

$$A = b_o'^2 + a_o^2 v_2 (m_{11}v_1 - 1)/\{2b_o m_{11}\} - (v_2/2b_o)\log(|V'|/|V|) \tag{35}$$

A boundary described by (34) is a parabola with vertex at the point

$$(a_o/(v_1 m_{11}),\ A - (v_2/b_o)\log\alpha) \tag{36}$$

and symmetric about the point

$$X_1 = a_o/v_1 m_{11} \tag{37}$$

When there are three populations with means given by (21) and $V = \text{diag}\{v_1, v_2\}$, we then have the regions $B_{3s}(s=1,2)$ bounded by a parabola $(\chi_1 - f)^2 = 4k(\chi_2 - g)$, where

$$f = a_o(1+a^2)/a^2 v_1$$
$$k = b_o(1+a^2)/2v_2 a^2 \tag{38}$$

and

$$g = b_o/2 + a_o^2\, {}_2(a^2 v_1 - 1 - a^2)/(2a^2 b_o) - (v_2/2b_o)\{\log(1+a^2/v_1) - 2\log\alpha\}$$

α being equal to 2 for B_{3a} and to 1 for B_{3b}.

5. Classification Probabilities

The acceptance region of Π_o being given by B_{is} (i=1,2,3 and s = a,b) under different assumptions, the true probability of misclassification of an element from the 'other' population Π into Π_o is given by

$$P(B_{is}|\Pi) = \sum_{j=1}^{m} [q_j/(1-q_o)] \int_{B_{is}} N_p(\mu_j, V) \tag{39}$$

and that of an element of Π_o into Π by

$$P(B_{is}^c|\Pi_o) = \int_{B_{is}^c} N_p(\mu_o, V), \tag{40}$$

where A^c denotes the complement of a set A. As the regions B_{2a}, B_{2b}, B_{3a} or B_{3b} are bounded by conicoids, the integrals in (39) and (40) can not be evaluated in closed form.

In special cases, when the populations are two dimensional and the common dispersion matrix is diagonal, the boundaries of B_{3a} and B_{3b} are parabolas. The integrals in (39) and (40) cannot be evaluated in closed form even in this case. Therefore, it is not possible to state precisely

in terms of misclassification probability how much we have to pay if we want to reduce our many population classification problem into a two population classification problem and which one of the alternative assumptions 1b, 2a, 2b, 3a and 3b is preferable next to the true assumption 1a.

In special cases, it may be possible to make some inference looking at the graphs.

Examples

In the following examples we consider three populations with densities $N_2(\mu_i, V)$, $i=0, 1, 2$, where

$$V = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}, \quad \mu_o = \begin{bmatrix} a_o \\ b_o \end{bmatrix}, \quad \mu_1 = \begin{bmatrix} -a \\ 0 \end{bmatrix} \text{ and } \mu_2 = \begin{bmatrix} a \\ 0 \end{bmatrix},$$

a, a_o, b_o being specified separately in each example. All the populations have the same prior probability.

Example 1. Let $a=1$, $a_o=3$ and $b_o=6$. Then we have

$$B_{1a} = \{(x_1y): 8x + 3y - 17 \geq 0, 4x + 3y - 17 \geq 0\},$$

$$B_{1b} = \{(x_1y): 8x + 3y - 15.b2 \geq 0, 4x + 3y - 15.b2 \geq 0\},$$

$$B_{2a} = \{(x_1y): 3y \geq 2 \log [1 + \exp (2x)] - 8x + 17\},$$

$$B_{2b} = \{(x_1y): 3y \geq 2 \log [1 + \exp (2x) - 8x + 15.b2\},$$

$$B_{3a} = \{(x_1y): (x-b)^2 \leq b(y-0.23)\},$$

$$B_{3b} = \{(x_1y): (x-b)^2 \leq b(y + 0.23)\}.$$

The classification procedure defining the partition (B_{1a}, B_{1a}^c) is optimal in the sense that it has the minimum expected cost of misclassification (which in this case is equal to the total misclassification probability). From the graph it is evident that the next best partition is (B_{2a}, B_{2a}^c), then (B_{2b}, B_{2b}^c), then (B_{1b}, B_{1b}^c). The two classification procedures based on the assumption of normality of the mixture perform relatively poorly.

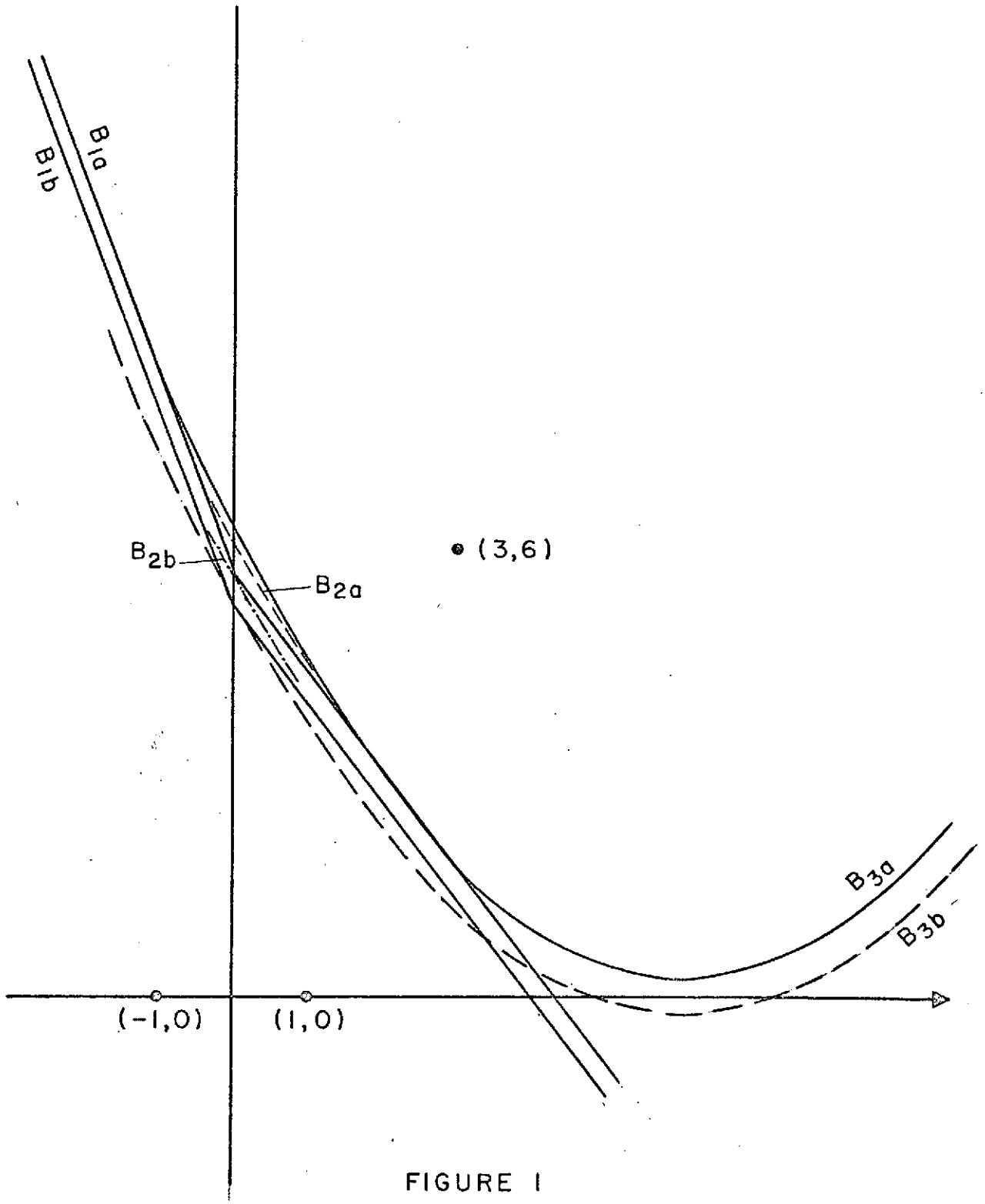


FIGURE 1

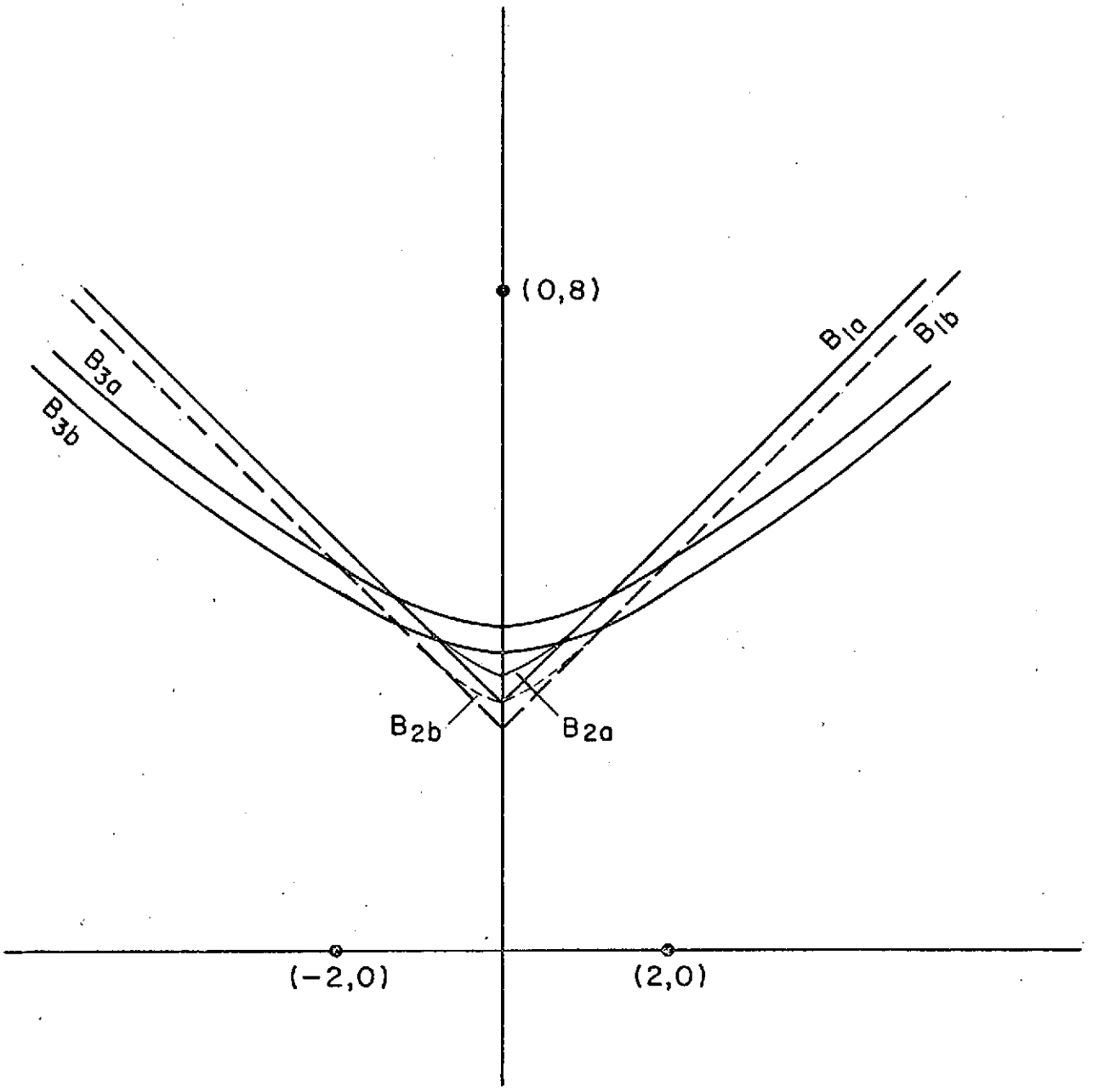


FIGURE 2

Of the two, the partition (B_{3a}, B_{3b}^c) has lower misclassification probability than the partition (B_{3b}, B_{3b}^c).

Example 2. Let $a=2$, $a_o=0$ and $b_o=8$. Then we have

$$B_{1a} = \{(x_1y): x + y - 3 \geq 0,\ y - x - 3 \geq 0\},$$

$$B_{1b} = \{(x_1y): x + y - 2.66 \geq 0,\ y - x - 2.66 \geq 0\},$$

$$B_{2a} = \{(x_1y): 2y \geq \log [1 + \exp (4x)] - 2x + 6\},$$

$$B_{2b} = \{(x_1y): 2y \geq \log [1 + \exp (4x)] - 2x + 5.31\},$$

$$B_{3a} = \{(x_1y): x^2 < 5(y - 3.94)\}$$

$$B_{3b} = \{(x_1y): x^2 < 5(y - 3.6)\}.$$

The classification procedure defining the partition (B_{1a}, B_{1a}^c) is optimal and from the graph we may infer that the next best is the one with partition (B_{2a}, B_{2a}^c). Unfortunately, in this case, no comparison of other procedures can be made even from the graph.

Reference

1. Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis, Wiley, New York.

2. Rao, C. R. (1965). Linear Statistical Inference and Its Applications, Wiley, New York.

EFFECT OF DISTANCE-MEASURES ON CLUSTER-BASED CLASSIFICATION PROCEDURES

P. L. ODELL and J. P. BASU

The University of Texas at Dallas

ABSTRACT

When we have a single sample of observations from two normal populations with same covariance matrix, but no training sample from either population, then for classifying future observations, the usual practice is to cluster the sample into two nearest neighbor clusters and design a Bayes' classifier treating the two clusters as two training samples. The use of ℓ_1-distance is often advocated for such clustering. It has been shown in this paper that such advocation is not always reasonable.

1. Introduction

We suppose that our data have been obtained through remote sensing devices from two p-variate normal populations Π_1 and Π_2 with densities $N_p(\mu_1,\Sigma)$ and $N_p(\mu_2,\Sigma)$ respectively, the vectors μ_1 and μ_2 of means and the common dispersion matrix Σ being unknown, and that we do not have any training sample, that is, we do not know the actual source of our observations; but the data is assumed to be such that two and only two distinct modes can be determined from them. For classifying future observations into these two populations on the basis of such data, the usual practice is to cluster the data into two nearest neighbor clusters C_1 and C_2 using metric or distance function defined by ℓ_1-, ℓ_2-, or ℓ_∞-norm; and then design a Bayes' classifier with the assumption that the population densities are $N_p(\hat{\mu}_j, \hat{\Sigma}_j)$ $(j=1,2)$, where $\hat{\mu}_j$ and $\hat{\Sigma}_j$ are respectively the sample mean and the sample dispersion matrix of the sample C_j assumed to be consisting of observations from the population Π_j alone. If it is known that the parent populations have the same dispersion matrix, then in designing the classifier, the matrices $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ above are replaced by $\hat{\Sigma}$, where

$$\hat{\Sigma} = [(N_1-1)\hat{\Sigma}_1 + (N_2-1)\hat{\Sigma}_2]/(N_1+N_2-2), \qquad (1)$$

N_j denoting the number of elements of C_j. A classification procedure defined in this fashion will be referred to as cluster based classification procedure.

Let P_o denote the Bayes' procedure based on complete knowledge of $N_p(\mu_j,\Sigma)$ $(j=1,2)$ and P_i the cluster based procedure when ℓ_i-norm $(i=1,2,\infty)$ has been used for clustering and the sample sizes are very large. Our objective in this paper is to compare the misclassification probabilities of these classification procedures. We have made the comparison a little simpler by taking $p=2$, $\mu_1=0$ and $\Sigma = I$, that is, by assuming our population densities to be $N_2(0,I)$ and $N_2(\mu_2,I)$. In designing the classifiers we have also assumed that both populations have equal prior probabilities (1/2 each) and equal misclassification costs. Now, without any further loss of generality, it can be assumed that $\mu_2^T = [a,b]$ and $a\geq 0$, $b\geq 0$.

2. Classification Procedure P_o

Under the set of assumptions made above, the Bayes' classifier is given by

$$W_o(Z) = ax + by - (a^2+b^2)/2, \quad Z^T = [x,y] \qquad (2)$$

and hence the acceptance region B_o for Π_1 is given by

$$B_o = \{(x,y): \; ax + by < (a^2+b^2)/2\}. \qquad (3)$$

If $P(j|i)$ denotes the probability of misclassifying an element from Π_i into Π_j, then for this procedure we have

$$P(2|1) = P(1|2) = \Phi(-\sqrt{a^2+b^2}/2), \qquad (4)$$

where

$$\Phi(t) = \int_{-\infty}^{t} \exp(-x^2/2)\, dx/\sqrt{2\pi}.$$

If a=1 and b=2, then we have from (4)

$$P(2|1) = P(1|2) = \Phi(-\sqrt{5}/2) = 0.13.$$

3. Cluster Based Classification

It is well known (Lerman, 1970) that as the sample sizes increase, the clusters C_{1i} and C_{2i} based on ℓ_i-norm converges with probability one to the partition (S_{1i}, S_{2i}) of R_2, the two dimensional real vector space, where

$$S_{2i} = S_{1i}^c = \{z = (x,y): \ ||z||_i < ||u_2 - z||_i\} \tag{5}$$

A^c denotes the complement of the set A and $||u-v||_i$ denotes the distance between the points u and v of R_2 according to ℓ_i-norm. $\hat{\mu}_{ji}$ and $\hat{\Sigma}_{ji}$, the mean and dispersion matrix of the sample C_{ji}, (j=1,2), can be viewed as the sample mean and sample dispersion matrix of a population with density function $p_{ji}(z)$ given by

$$p_{ji}(z) = (1/2)\{N_2(0,I) + N_2(\mu_2,I)\}, \ z \in S_{ji} \tag{6}$$
$$= 0 \text{ otherwise.}$$

Thus, by the law of large numbers, with probability one $\hat{\mu}_{ji}$ converges to μ_{ji} and $\hat{\Sigma}_{ji}$ to Σ_{ji}, where

$$\mu_{ji} = \int_{S_{ji}} z p_{ji}(z)dz / P(S_{ji}) \tag{7}$$

$$\Sigma_{ji} = \int_{S_{ji}} (z-\mu_{ji})(z-\mu_{ji})^T p_{ji}(z)dz / P(S_{ji}) \tag{8}$$

and $$P(S_{ji}) = \int_{S_{ji}} p_{ji}(z)dz. \tag{9}$$

In designing cluster based classifiers, as we have mentioned earlier, we shall consider two cases--when we do not know that the dispersion matrices of the population Π_1 and Π_2 are equal and when we know that they are equal. In the first case, we shall denote the cluster based procedure by P_j (based on ℓ_i-norm) and in the second case by P_j'. Thus P_i' is the Bayes' procedure based on the densities $N_2(\mu_{ji}, \Sigma_i)$ (j=1,2), where

$$\Sigma_i = P(S_{ji})\Sigma_{1i} + P(S_{2i})\Sigma_{2i}. \tag{10}$$

Misclassification Probability

If a classification procedure defines a set B as the acceptance region for the population Π_1 with true density $N_2(0,I)$ and B^c as the acceptance region for Π_2 with true density $N_2(\mu_2,I)$, then the misclassification probabilities of this procedure are given by

$$P(2|1) = P(B^c|\Pi_1) = \int_{B^c} N_2(0,I)dz$$

and $$P(1|2) = P(B|\Pi_2) = \int_B N_2(\mu_2,I)dz \tag{11}$$

Let B be the acceptance region for Π_1 according to the Bayes' classification procedure in which the population densities for Π_1 and Π_2 have been inadvertently assumed to be N (m,V) and N (m',V') respectively. Then

$$B = \{Z = (x,y): \ W(x,y) \ \ 0\}, \tag{12}$$

where

$$W(x,y) = 2(g_1x+g_2y) + K + 2exy - d_1x^2 - d_2y^2, \tag{13}$$
$$d_1 = v'_{22}/|V'| - v_{22}/|V|,$$
$$d_2 = v'_{11}/|V'| - v_{11}/|V|,$$
$$e = v'_{12}/|V'| - v_{12}/|V|,$$
$$g_1 = (v'_{22}m'_1 - v'_{12}m'_2)/|V'| - (v_{22}m_1 - v_{12}m_2)/|V|,$$
$$g_2 = (v'_{11}m'_2 - v'_{12}m'_1)/|V'| - (v_{11}m_2 - v_{12}m_1)/|V|,$$
$$K = m^T V^{-1} m - m'^T V'^{-1} m' - \log(|V'|/|V|).$$

Here $V = [v_{ij}]$, $V' = [v'_{ij}]$, $m^T = [m_1, m_2]$ and $m'^T = [m'_1, m'_2]$.

The misclassification probabilities P(2|1) and P(1|2) can now be expressed in the form

$$P(2|1) = P(W(X,Y) \geq 0 \ | Z \in \Pi_1) \tag{14}$$

and $$P(1|2) = P(W(X,Y) < 0 \ | Z \in \Pi_2) \tag{15}$$

If $V \neq V'$, then the exact distribution of W(X,Y) when $Z \in \Pi_1$ or $Z \in \Pi_2$ is intractable. But if d_1, d_2 and e are negligible compared to g_1 and g_2, then in (14) and (15), the contribution from the second degree terms in X and Y will be negligible compared to that from $K + 2(g_1X+g_2Y)$. Thus we can perhaps approximate the distribution of W(X,Y) by the distribution of $\hat{W}(X,Y) = K + 2(g_1X+g_2Y)$. But if V=V', then $W=\hat{W}$. Now $\hat{W}$ can be shown to be distributed normally with mean

$$E\hat{W} = K, \text{ when } Z \sim N_2(0,I)$$

and $$E\hat{W} = K + 2(g_1a+g_2b), \text{ when } Z \sim N_2(\mu_2,I), \mu_2^T = [a,b] \tag{16}$$

and variance

$$\text{Var } \hat{W} = 4(g_1^2+g_2^2) \tag{17}$$

in either case. The approximate values of the misclassification probabilities P(2|1) and P(1|2) can be given by

$$P(2|1) \doteq P(\hat{W}>0 | Z \in \Pi_1) = 1-\Phi(-K/2\sqrt{g_1^2+g_2^2}) \tag{18}$$

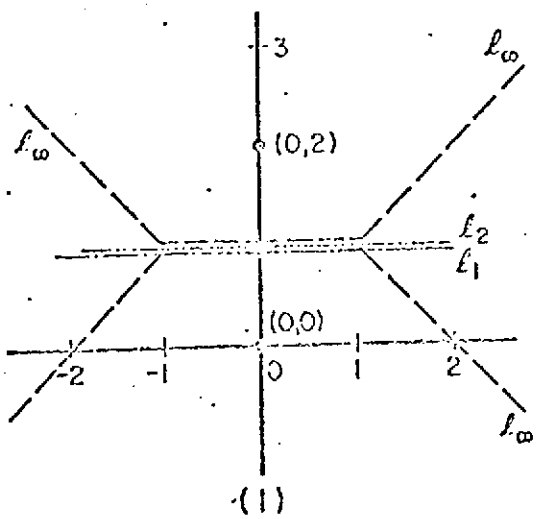
and $$P(1|2) \doteq P(\hat{W}\leq 0 | Z \in \Pi_2) = \Phi(-\{2(g_1a+g_2b)+K\}/2\sqrt{g_1^2+g_2^2}). \tag{19}$$

4. Classification Procedure P_2 and P_2'

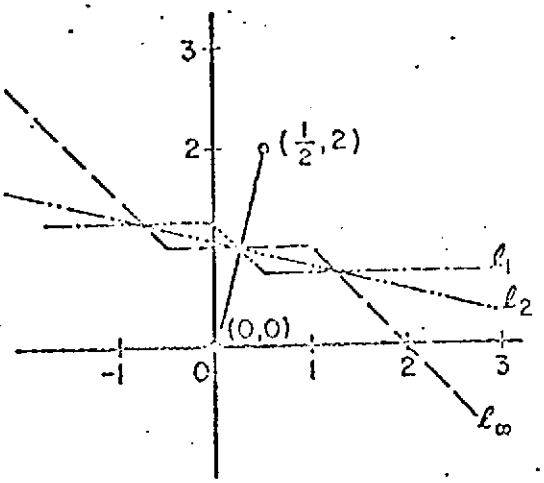
The ℓ_2-norm induced distance between two points $u = [u_1,u_2]^T$ and $v = [v_1,v_2]^T$ is given by

$$||u-v||_2 = (u_1-v_1)^2 + (u_2-v_2)^2)^{1/2}.$$

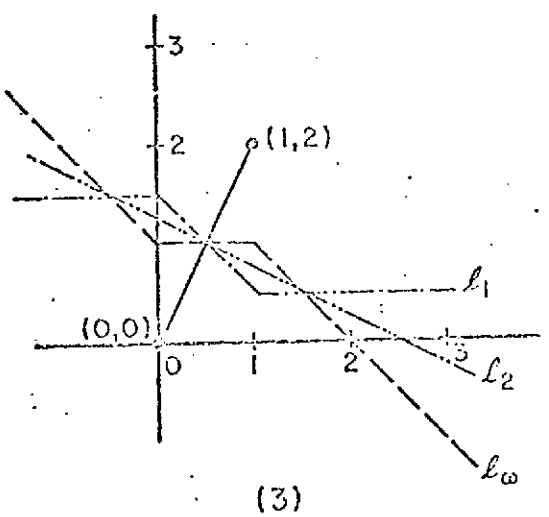
Since $\mu_2^T = [a,b]$, then it follows from (5) that



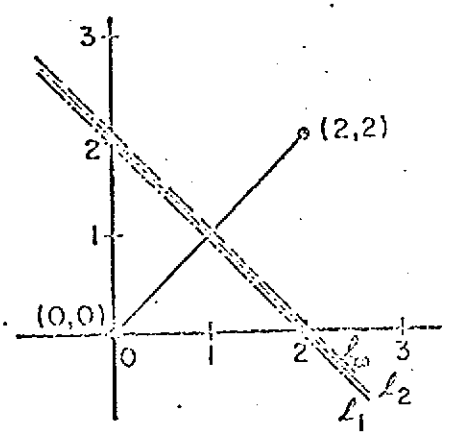
(1)



(2)

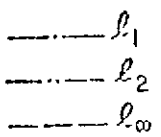


(3)



(4)

Limiting clusters using l_1-, l_2- and l_∞-norm



the limiting clusters S_{12} and S_{22} are given by

$$S_{22}^c = S_{12} = \{(x,y): ax+by < (a^2+b^2)/2\}.$$

Evidently $S_{12}=B_o$, as given by (3). Now it follows from (6)-(9) that

$$P(S_{12}) = P(S_{22}) = 1/2, \tag{20}$$

$$\mu_{i2} = k_i\begin{bmatrix}a\\b\end{bmatrix} \text{ and}$$

$$\Sigma_{i2} = \begin{bmatrix}1 + a^2k_i(1-k_i) & abk_i(1-k_i)\\ abk_i(1-k_i) & 1 + b^2k_i(1-k_i)\end{bmatrix}, \tag{21}$$

(i=1,2), where

$$k_1 = \Phi(-a^2+b^2/2) - (2\Pi(a^2+b^2)/2)^{-1} \exp\{-(a^2+b^2)/8\} \tag{22}$$

and $k_2 = \Phi(a^2+b^2/2) + (2\Pi(a^2+b^2)/2)^{-1} \exp\{-(a^2+b^2)/8\}.$ (23)

The set B_2, the acceptance region of Π_1 according to P_2, can be obtained from (12) by letting $m=\mu_{12}$, $m'=\mu_{22}$, $V=\Sigma_{12}$ and $V'=\Sigma_{22}$ and the corresponding misclassification probabilities from (13), (18) and (19).

An Example

Let a=1 and b=2. Then

$$S_{22}^c = S_{12} = \{(x,y): x+2y \le 2.5\}$$

[figure (3)]

and it follows from (21), (22) and (23) that

$$k_1 = -0.06,\ k_2 = 1.06,\ \mu_{i2}^T = k_i[1,2],$$

$$\Sigma_{12} = \begin{bmatrix}0.94 & -0.11\\ -0.11 & 0.78\end{bmatrix} \text{ and } \Sigma_{22} = \begin{bmatrix}0.93 & -0.12\\ -0.12 & 0.75\end{bmatrix}.$$

In P_2, for discriminating observations from Π_1 and Π_2, the Bayes' classifier based on densities $N_2(\mu_{12},\Sigma_{12})$ and $N_2(\mu_{22},\Sigma_{22})$ is used instead of that based on the true densities $N_2(0,I)$ and $N_2([1,2]^T,I)$, which are unknown. Then the approximate values of the resulting misclassification probabilities are obtained from (18) and (19) as

$$P(1|2) = 0.22 \text{ and } P(2|1) = 0.078.$$

In procedure P_2', the Bayes' classifier is based on densities $N_2(\mu_{12},\Sigma)$ and $N_2(\mu_{22},\Sigma)$, where

$$\Sigma = P(S_{12})\Sigma_{12} + P(S_{22})\Sigma_{22} = \begin{bmatrix}0.935 & -0.115\\ -0.115 & 0.765\end{bmatrix}.$$

The misclassification probabilities are now obtained from (18) and (19) as

$$P(1|2) = 0.18 \text{ and } P(2|1) = 0.09.$$

5. Classification Procedure P_1 and P_1'

The ℓ_1-norm induced distance between the points u and v is given by

$$||u-v||_1 = |u_1-v_1| + |u_2-v_2|.$$

The limiting cluster $S_{11}(=S_{21}^c)$ consists of points (x,y) satisfying the following conditions in three different cases.

Case 1. b>a>0 [figure (2) and (3)]

(i) $x \ge a$, $y<(b-a)/2$

(ii) $0<x<a$, $x+y<(a+b)/2$

(iii) $x<0$, $y<(a+b)/2$.

In this case S_{11} and S_{21} are distinct from S_{12} and S_{22} given by (3).

Case 2. b>0, a=0 [figure (1)]

(i) $y<b/2$.

In this case S_{11} coincides with S_{12} and S_{21} coincides with S_{22}.

Case 3. b=a>0 [figure (4)]

(i) $x+y<a$

The ℓ_1-norm distances of the point (x,y) from (0,0) and (a,b) are equal when

$x+y \ge a$ and $x<0$

or $x+y \ge a$ and $y<0$.

These points have been assigned to S_{21} in order that no point escape classification. In this case also the cluster S_{11} coincides with S_{12} and S_{21} coincides with S_{22}.

Thus the procedure $P_1(P_1')$ differs from the procedure $P_2(P_2')$ only in case 1. In this case, the expression for the components of the means μ_{11} and μ_{21} and the dispersion matrices Σ_{21} and Σ_{21} are long and complicated expressions involving the probabilities $\phi(a)$, $\phi[(a+b)/2]$ and $\phi[(b-a)/2]$. For that reason we have not included it here. As in the example in section 4, if we take a=1 and b=2, then we shall have

$$\mu_{11} = \begin{bmatrix}-0.006\\ 0.197\end{bmatrix}, \qquad \mu_{21} = \begin{bmatrix}0.423\\ 2.11\end{bmatrix}$$

$$\Sigma_{11} = \begin{bmatrix}0.98 & -0.06\\ -0.05 & 0.66\end{bmatrix} \text{ and } \Sigma_{21} = \begin{bmatrix}1.82 & -.2\\ -.2 & 0.8\end{bmatrix}$$

From (13) it now follows that

$$W(X,Y) = 0.46X^2 + 0.24Y^2 + 2(0.52X + 2.46Y) + 0.04XY - 9.36$$

The exact distribution of W is intractable. The normal approximation to W given in section 3 is not valid, since the coefficients 0.46 of X^2 and 0.24 of Y^2 are not negligible compared to coefficients 0.52 of X and 2.46 of Y. Using a very rough approximation, it can be shown that in this case $P(2|1)$ and $P(1|2)$ are larger than those for P_2.

For procedure P_1' we have

$$\Sigma_1 = P(S_{11})\Sigma_{11} + P(S_{21})\Sigma_{21}$$

$$= 0.495\,\Sigma_{11} + 0.505\,\Sigma_{21}$$

$$= \begin{bmatrix} 1.4 & -0.13 \\ -0.13 & 0.73 \end{bmatrix}$$

From (13) we now have

$$W(X,Y) = 1.11SX + 5.44Y - 6.541$$

and from (18) and (19) we have

$$P(2|1) = 0.18 \text{ and } P(1|2) = 0.12$$

6. Classification Procedure P_∞ and P'_∞

The ℓ_∞-norm induced distance between the points u and v is given by

$$||u-v||_\infty = \max\{|u_1-v_1|, |u_2-v_2|\}.$$

The points (x,y) in the limiting cluster $S_{1_\infty}(=S_{2_\infty})$ satisfy the following conditions.

Case 1. b>a>0 [figure (2) and (3)]

(i) $a-b/2<x<\max(b/2,a)$ and $y<b/2$,

(ii) $x<a-b/2$ and $x+y<a$

(iii) $x>\max(b/2,a)$ and $x+y<b$.

Case 2. b=a>0 [figure (4)]

(i) $x+y<a$

The ℓ_∞- distances of a point (x,y) from (0,0) and (a,b) are equal when

$x+y\geq a$ and $x<0$

or $x+y\geq a$ and $y<0$.

Such points have been assigned to S_{2_∞} arbitrarily. Thus in this case S_{1_∞} and S_{2_∞} coincide with S_{12} and S_{22} respectively.

Case 3. b>0, a=0 [figure (1)]

In this case, the performance of ℓ_∞-norm in clustering is very poor. The ℓ_∞- distances of every point (x,y) from (0,0) and (0,b) are equal whenever

$x-y+2<0$ and $x+y<0$

or, $x+y-2>0$ and $x-y>0$

However, we can arbitrarily define S_{1_∞} as the set

$\{(x,y): y<b/2\}$.

But S_{1_∞} then becomes identical to S_{11} or S_{12}.

Thus the procedure $P_\infty(P'_\infty)$ differs from $P_2(P'_2)$ only in case 1. But in this case, the exact misclassification probability is hard to find. In the special case, when a=1 and b=2, (figure (3)) we find that the boundary line of S_{1_∞} and S_{2_∞} can be obtained as the reflection of the boundary line of S_{11} and S_{21} about the boundary line of S_{12} and S_{22}. From this we may guess that, as in case of the procedure P_1 and P'_1, the total misclassification probabilities of P_∞ and P'_∞ will be larger than the total misclassification probabilities of P_2 and P'_2.

7. Conclusion

Our findings about the misclassification probabilities $P(1|2)$ and $P(2|1)$ for the procedures P_0, P_1, P'_1, P_2 and P'_2 in the special case when a=1 and b=2 can be seen at a glance from the following table.

Table

	True Bayes	ℓ_1-norm based		ℓ_2-norm based	
	P_0	P_1	P'_1	P_2	P'_2
$P(1\|2)$	0.13		0.12	0.22	0.18
$P(2\|1)$	0.13		0.18	0.078	0.09
Total	0.26	*	0.30	0.299	0.27

*Using rough approximation it has been found that the total misclassification probability of P_1 is greater than 0.229.

Thus, taking misclassification probability as the criterion for comparing the performance of the classifiers, we find that in the example considered above P'_2 is better than P'_1 and P_2 is better than P_1. But, no such conclusion can be made in general. It is not known whether P'_2 always has lower misclassification probability than its competitor P'_1 or P'_∞ and P_2 has lower misclassification probability than P_1 or P_∞.

Among all norm-induced distances, ℓ_1-distances are the most easy to compute. When population means are known to be such that the straight line joining them is inclined at an angle of 45° to x-axis (for example, when they are (0,0) and (a,a), the classification procedure $P_1(P'_1)$ and $P_2(P'_2)$ become identical. Only in that case, P_1 or P'_1 is the optimal procedure, because we then achieve the computational ease along with low misclassification probability.

The procedures P_∞ and P'_∞ are not desirable. If in the process of clustering, two cluster centers are such that they have the same x-coordinate, then some points are to be clustered according to ℓ_2-distance. This arbitrariness in allocation of points to the clusters makes P_∞ or P'_∞ undesirable.

Reference

1. Lerman, I. C. (1970). Les Bases de la Classification Automatique. Gauthier-Villars, Paris.

ADAPTIVE PATTERN RECOGNITION: A SURVEY

J. P. Basu and P. L. Odell

THE UNIVERSITY OF TEXAS AT DALLAS

This research has been supported by Johnson Space Research Center Contract # NAS9-13512

1. Introduction

A classifier or discriminant function W(xjc) is a function or a set of functions which determine the membership of an observation vector x into one of several given sub-populations, c denoting the vector of parameters of the classifier. The classifier usually considered is one which has minimum expected cost of misclassification among all classifiers. This optimal classifier is known as Bayes' classifier [2]. This classifier is obtained on the assumption of complete knowledge of misclassification costs, prior probabilities (the proportion of the given subpopulations in the over all population) and the probability densities of the given subpopulations. When the parametric family of the population densities are known, but the parameter themselves are unknown, then the usual practice is to replace the true values of these parameters by their respective estimates based on sets of past observations of known classification in the expressions giving the elements of c. A classifier $W(x, \hat{c})$, thus obtained, lacks in the above optimality property. A set of past observations known to be coming from a particular population is often referred to as a training sample. When the parametric family of the population densities are also unknown and sometimes the prior probabilities are also unknown, the methods by which a classifier W(x, c) is obtained from the available training samples of the populations are referred to as nonparametric methods of classifier design. An adaptive pattern recognition scheme is one such nonparametric method.

An adaptive pattern recognition scheme is a nonparametric method in which a classifier is assumed to be of the form

$$W(x, c) = c_1 Q_1(x) + \ldots + c_m Q_m(x) = c^T \Phi(x), \tag{1}$$

where the components $Q_1(x), \ldots, Q_m(x)$ of the vector $\Phi(x)$ are prechosen linearly independent continuous functions of the observation vector x and the unknown parameters $[c_1, \ldots, c_m] = c^T$ are directly estimated recursively as in [1, 7] or nonrecursively as in [6, 8] from the available training samples, instead of being obtained terms of the estimates of the probability densities. A classifier obtained in this fashion will be referred to as an adaptive classifier.

The functional forms of the functions $Q_i(x)$'s and their number m are often choosen arbitrarily. This arbitrariness in the choice of the functions $Q_i(x)$'s may create problem. A good classifier should be optimal for discriminating among the given populations. For example, it is well known [2] that when there are two p-dimensional normal populations with different covariance matrices, a linear classifier of the form

$$W(x, c) = c_1 x_1 + \ldots + c_p x_p + c_o = c^T z, \tag{2}$$

where $c^T = [c_1, \ldots, c_p, c_o]$, $x^T = [x_1, \ldots, x_p]$ and $z^T = [x^T, 1]$, is not optimal for discriminating between these two populations. Therefore, if the functional forms of $Q_i(x)$'s are not properly chosen, then the classifier $W(x, c)$ given by (1) may not be optimal for discriminating among the given populations. Also in designing an adaptive classifier rarely any attention is paid to the optimality criterion, minimum expected cost of misclassification. However, adaptive classifiers are easy to design and are very attractive in fast classification for their simplicity. Besides, there are empirical cases in which they do no worse, if not better, than many other sample based classifiers [2].

A comprehensive survey of adaptive pattern recognition schemes can be found in H_0 and Agrawala [5] and in Tsypkin [9]. Our objective here is to present a unified theory of adaptive pattern recognition. Most of the adaptive pattern recognition schemes have been devised for discriminating between two populations and few of these schemes have been generalized in a very straight forward manner for discriminating among more than two populations. For this reason, we shall also restrict ourselves mostly to the case of two population classification.

2. Basic Theory

Let all the observations come from two p-dimensional population π_1 and π_2. Then the observations can be represented by points and the two training samples by two sets T_1 and T_2 of points in the p-dimensional real Euclidean space E_p. It will be evident later from the way an adaptive classifier is designed that the sets T_1 and T_2 should be disjoint. A real function $f(x)$, $f: E_p \to E_1$, is said to separate two disjoint sets T_1 and T_2 in E_p, if there exists a constant t such that $f(x) > t$ for all $x \in T_1$ and $f(x) < t$ for all $x \in T_2$. If there exists a function $f(x)$ separating T_1 and T_2, then $f(x)$ can be considered to be an ideal classifier which can be used to classify all of the points of T_1 and T_2 correctly. It may not often be possible to construct such $f(x)$, even if it exists. But, it may be possible to obtain a function $W(x, c)$ as an approximation to the function $f(x)$ and use it as a classifier instead of the ideal classifier $f(x)$. This is precisely the way an adaptive classifier is designed.

The existence of function $f(x)$ separating the training samples T_1 and T_2 will be evident soon. The following theorem is well known (Dunford and Schwartz, [4], p. 417).

Theorem. In a topological vector space, any two disjoint convex sets, one of which has an interior point, can be separated by a nonzero continuous linear functional.

If T_1 and T_2 have disjoint convex hulls C_1 and C_2, then we can use the above theorem to prove the existence of function f(x) of the form

$$f(x) = c_1x_1 + \ldots + c_px_p + c_o$$

such that $f(x) > 0$ on C_1 and $f(x) < 0$ on C_2. In this case the training samples T_1 and T_2 are said to be linearly separable. Therefore, when the training samples are linearly separable, then a linear function f(x) separating T_1 and T_2 can be used as an ideal classifier. An adaptive classifier is then easily obtained as a linear function approximating this ideal classifier.

Unfortunately, the sets T_1 and T_2 of training samples are rarely linearly separable. Therefore, often there may not exist any linear function separating T_1 and T_2. However, there exists a function f(x), not necessarily linear, which will separate T_1 and T_2 whenever the sets C_1 and C_2, the smallest closed sets containing T_1 and T_2 respectively, satisfy the conditions of Uryshon's theorem, well known in Topology ([4], p. 24).

Uryshon's Theorem. If C_1 and C_2 are disjoint subsets of a normal topological space V, then there exists a continuous function g(x), $g: V \to [0, 1]$, such that $g(x) = 0$ for all $x \in C_2$ and $g(x) = 1$ on C_1.

Using the function g(x) of Uryshon's theorem, we can easily define a continuous function f(x) on V such that (a) $f(x) = t_1$ on C_1 and $f(x) = t_2$ on C_2, where t_1 and t_2 are any two distinct preassigned numbers, or, (b) $f(x) > \delta$ on C_1 and $f(x) < -\delta$ on C_2, where δ is a preassigned positive number. Now, since

it is well known that E_p is a normal topological space under any nontrivial topology and T_1 and T_2 are disjoint sets containing finite number of points, then we can always obtain two disjoint sets C_1 and C_2 as the smallest closed sets containing T_1 and T_2. Therefore, as long as T_1 and T_2 are disjoint sets of training samples, there always exists a function f(x), linear or nonlinear, which separates T_1 and T_2.

As we have mentioned earlier, in an adaptive pattern recognition procedure the ideal classifier f(x) is approximated by a function W(x, c) of the form (2), namely, a linear function of an observation x to obtain a linear adaptive classifier, or, by a function W(x, c) of the form (1), a linear combination of a finite number of known nonlinear continuous functions, to obtain a nonlinear adaptive classifier, the parameter vector c being selected in order to minimize certain given pay-off function.

As the probability densities of the populations are unknown, expected misclassification cost cannot be evaluated and therefore cannot be used as a pay-off function. Distances D(f, W) between the functions f(x) and W(x, c) or some functions of it defined by some meaningful metric can be used as geometrically and intuitively appealing pay-off functions. For example, using Euclidean metric we can define

$$D_2(f, W) = \sum_{x \in T_1 \cup T_2} | f(x) - W(x, c) |^2, \tag{3}$$

as a pay-off function and select the parameter vector c in order to minimize D_2 (f, W). In this case W(x, c) is a least square approximation to f(x). In order to assure the existence of a unique minimum of the pay-off function, a pay-off function of the form

$$L(f, W) = \sum_{x \in T_1 \cup T_2} F\,(f(x) - W(x, c))/n, \tag{4}$$

where n is the total number of points in T_1 and T_2 and in the fashion of decision theory F is a convex function interpreted as a loss function, can also be used [9]. We may note that the pay-off function $\bar{D}_2\,(f, W) = D_2(f, W)/n$, often referred to as mean square error criterion [6, 8], is a particular case of (4).

Let the subpopulation Π_i have prior probability q_i and probability density function $p_i(x)$. Then the population which is a mixture of Π_1 and Π_2 has probability density function $q_1 p_1(x) + q_2 p_2(x)$. Therefore, the expected value of $F(f(x) - W(X, c))$ will be given by

$$\begin{aligned} R(f, W) &= E\,F\,(f(x) - W(X, c)) \\ &= \sum_{i=1}^{2} q_i \int F(f(x) - W(x, c))\; p_i(x)\; dx \end{aligned} \tag{5}$$

Now, if there are n_i observations in T_i and $n_1 + n_2 = n$, then we can write

$$L(f, W) = \sum_{i=1}^{2} (n_i/n) \sum_{x \in T_i} F(f(x) - W(x, c))/n_i \tag{6}$$

and observe that L(f, W) is the sample mean and R(f, W) is the population mean of $F\,(f(x) - W(X, c))$.

It is important to note that there is some arbitrariness in the choice of the number m and the continuous functions $Q_1(x), \ldots, Q_m(x)$, used in the definition of W(x, c).

3. Linear Classifier

The linear classifiers are used in adaptive classification mainly due to following reasons.

1. They are relatively simple to implement as electronic circuits.

2. The circuits can be made adaptive very easily.

3. The Bayes' classifier with minimum expected cost of misclassification has a linear structure when the two populations have normal distribution with equal covariance matrices.

3.1. Nilsson's Classifier

The algorithm proposed by Nilsson [7] iteratively determines the parameter vector c of $W(x, c)$, as defined in (2), according to the following rule. Let $x(i)$ denote the ith observation of known classification, that is, an element selected from $T_1 \cup T_2$ at the ith step of an iteration, $z(i)$ the corresponding value of z, $c(i, n)$ that of c at the ith step of the nth iteration and $W_{in} = c^T(i, n)\, z(i)$. Then at the $(i+1)$th step of the nth iteration we have

$$
\begin{aligned}
c(i+1, n) &= c(i, n) && \text{if } x(i) \in T_1 \text{ and } W_{in} > 0 \\
&= c(i, n) && \text{if } x(i) \in T_2 \text{ and } W_{in} < 0 \\
&= c(i, n) + dz(i) && \text{if } x(i) \in T_1 \text{ and } W_{in} \leq 0 \\
&= c(i, n) - dz(i) && \text{if } x(i) \in T_2 \text{ and } W_{in} \geq 0,
\end{aligned}
$$

where d is a positive number, so chosen that $c^T(i+1, n)\, z(i)$ can correctly classify $x(i)$ for all i. The vector $c(1, n+1)$ is taken as the value of c at the end of nth iteration. The process is continued until we have, for some i, $c(i, n) = c(j, n) = c(\ell, n+1) = \hat{c}$ (say) for all $j>i$ and $\ell<i$. When such $\hat{c}$ exists, we shall say that $c(i, n)$ has converged to $\hat{c}$.

Unfortunately, if the sets T_1 and T_2 of training samples are not linearly separable, then the sequence c(i, n) may not converge. The algorithm is effective, that is, c(i, n) converges to a $\hat{c}$, only when T_1 and T_2 are linearly separable. It should be noted that $\hat{c}$ is not unique; it depends on the sequence in which we select the observations from $T_1 \cup T_2$. The number of iterations needed to reach such $\hat{c}$ also depends on this sequence. The classifier $W(x, \hat{c})$ constructed in this fashion is not in general a Bayes classifier.

Example. Let $T_1 = \{[0, 1]^T, [1, 1]^T\}$, $T_2 = \{[0, 0]^T, [1, 0]^T\}$. Then choosing d = 1 and the sequence of observations as in the Table we can obtain Nilsson's classifier in the following way.

		c			z				
n	i	c_1	c_2	c_o	x_1	x_2	1	W_i	Change c?
1	1	0	0	0	1	1	1	0	Yes
	2	1	1	1	1	0	1	2	Yes
	3	0	1	0	0	1	1	1	No
	4	0	1	0	0	0	1	0	Yes
2	1	0	1	-1	1	1	1	0	Yes
	2	1	2	0	1	0	1	1	Yes
	3	0	2	-1	0	1	1	1	No
	4	0	2	-1	0	0	1	-1	No
3	1	0	2	-1	1	1	1	1	No
	2	0	2	-1	1	0	1	-1	No

Therefore $\hat{c}^T = [0, 2, -1]$ and the classifier is $2x_2 - 1$. A decision rule based on this classifier is to assign x to Π_1 if $2x_2 - 1 > 0$ and to Π_2 if $2x_2 - 1 < 0$.

$$\ell = d\hat{\Sigma}^{-1}(m_2-m_1) \text{ and } c_o = -\{\ell^T(N_1m_1+N_2m_2) - (N_1-N_2)\}/(N_1+N_2), \quad (9)$$

where $$d = 2/[(N_1+N_2)^3/\{N_1N_2(N_1+N_2-2)\} + (m_1-m_2)^T\hat{\Sigma}^{-1}(m_1-m_2)] \quad (10)$$

and $$\hat{\Sigma} = (N_1S_1+N_2S_2)/(N_1+N_2-2). \quad (11)$$

When $N_1 = N_2$, then we have

$$c_o = (m_1-m_2)^T\hat{\Sigma}^{-1}(m_1+m_2)[Nd/4(n-1)] \quad (12)$$

The classifier $W(x, c) = c^Tz$ in this case resembles the Bayes' classifier with parameters of the population densities replaced by their respective estimates, that is, Anderson's [2] sample based discriminant function for two normal population with equal covariance matrices.

The classifier attributed to Agmon and Mays [5] is obtained when

$$L(f, W) = \sum_{x\varepsilon T_1 \cup T_2} [|W(x, c) - f(x)| - \{W(x, c) - f(x)\}]^2$$

is used as the pay-off function.

4. Nonlinear Classifier

The nonlinear classifiers are used in adaptive classification mainly due to following reasons.

1. If the sets T_1 and T_2 of training samples are not linearly separable, then a function $f(x)$ that can be used to separate them are often nonlinear.

2. The Bayes classifier has nonlinear structure when the two populations have (a) multivariate normal distributions with unequal covariance matrices, or (b) nonnormal multivariate distributions.

3.2 Other Linear Classifiers

An adaptive linear classifier can be obtained as linear approximation to a function f(x) separating T_1 and T_2 and taking the value 1 on T_1 and -1 on T_2, the parameter vector c being selected as a vector $\hat{c}$ that minimize a given pay-off function. Different pay-off functions can be used to obtain different classifiers.

Koford-Groner Classifier

An adaptive linear classifier W(x, c) may be obtained by selecting a vector $\hat{c}$ as the parameter vector that minimize the mean square error criterion $\bar{D}_2$ (f, W), or equivalently the squared Euclidean distance D_2 (f, W) given by (3), using gradient method or any other suitable optimization technique. An iterative algorithm based on gradient method for obtaining $\hat{c}$ can be given by

$$c(n+1) = c(n) + \rho z \sum_{x \in T_1 \cup T_2} \{f(x) - c^T z\} \qquad (7)$$

Koford and Groner [6] observed that the criterion $D_2(f, W)$ can be expre expressed in the form

$$D_2(f, W) = \sum_{j=1}^{2} N_j [\ell^T S_j \ell + \{\ell^T m_j + c_o + (-1)^j\}^2], \qquad (8)$$

where $C^T = [\ell^T, c_o]$, $m_j = \sum_{x \in T_j} x/N_j$, $S_j = \sum_{x \in T_j} (x-m_j)(x-m_j)^T/N_j$

and N_j is the number of observations in T_j. The values of ℓ and c_o that minimize D_2 (f, W), given by (8), are as follows.

As mentioned earlier, a nonlinear classifier $W(x, c)$ is obtained in the form of a linear combination

$$W(x, c) = c^T \Phi(x) = c_1 Q_1(x) + \ldots + c_m Q_m(x) \tag{1}$$

of linearly independent nonlinear continuous functions $Q_1(x), \ldots, Q_m(x)$, whose functional forms and number are prechosen. The performance of the classifier $W(x, c)$, how often the classifier will be able to correctly classify an observation of known classification, depends primarily on the choice of these functions. But, often they are chosen arbitrarily on the basis of economic consideration and one's intuition. However, sometimes, when one has some prior knowledge of the population densities, they can be satisfactorily chosen. For example, when it is known that the populations have multivariate normal distributions with unequal covariance matrices, the function $W(x, c)$ may be chosen as a second degree polynomial in $x_1, \ldots, x_p$, the p components of the observation vector x, and the functions $Q_i(x)$'s as the monomials $x_1^{k_1} x_2^{k_2} \ldots x_p^{k_p}$, where k_i's are 0, 1 or 2 and $k_1 + k_2 + \ldots + k_p = 2$. It has been suggested [5] that when the densities are unknown, the functions $Q_i(x)$'s should be taken as orthogonal or orthonormal functions such as generalized Hermite functions given by

$$H^{(n)}_{k_1, k_2, \ldots, k_p}(x) = (-1)^n (2\pi)^{-n/2} \frac{\partial^n}{\partial x_1^{k_1} \ldots \partial x_p^{k_p}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{p} x_i^2\right). \tag{13}$$

Unfortunately, there is no specific rule to guide us in selecting these functions judiciously.

4.1. Patterson — Womack Classifier

The adaptive classifier of Patterson and Womack [8] is a nonlinear classifier $W(x, c)$ obtained as an approximation to a function $f(x)$ that separates the training sample sets T_1 and T_2 and takes the value $C(2|1)$ on T_1 and $-C(1|2)$ on T_2, where $C(j|i)$ denotes the cost of misclassifying an element from Π_i into Π_j. The parameter vector c is selected as a vector $\hat{c}$ that minimize the pay-off function

$$M(N_1, N_2) = (q_1/N_1) \sum_{x \in T_1} |W(x, c) - C(2|1)|^2 +$$

$$(q_2/N_2) \sum_{x \in T_2} |W(x, c) + C(1|2)|^2, \tag{14}$$

where N_i is the number of elements in T_i and the prior probability q_i of the population Π_i is assumed to be known. As in (5) and (6), it is not difficult to see that $M(N_1, N_2)$ is the sample average of $|W(X, c) - f(X)|^2$ and therefore, by law of large number, $M(N_1, N_2)$ converges to

$$R(f, W) = E|W(X, c) - f(X)|^2$$

as $N_1 \to \infty$ and $N_2 \to \infty$. Patterson and Womack [8] have shown that the vector $\hat{c}$ that minimize $R(f, W)$, as given above, will also minimize

$$E\ |W(X, c) - D(X)|^2,$$

where $D(X)$ is the Bayes' classifier given by

$$D(X) = \{C(2|1)\ q_1 p_1(X) - C(1|2)\ q_2 p_2(X)\}/\{q_1 p_1(X) + q_2 p_2(X)\} \qquad (15)$$

and $p_i(x)$ is the density of Π_i, provided that $f(x)$ has been chosen in the above fashion. This has been the motivation behind using $C(2|1)$ and $-C(1|2)$ as the preassigned values of the separating function $f(x)$.

When q_i's are unknown and to be estimated, then $N_i/(N_1+N_2)$ can be used as an unbiased estimate of q_i, where N_i is the number of elements belonging to T_i out of a sample of size (N_1+N_2) from the mixed population. Then the criterion $M(N_1, N_2)$ may be replaced by

$$\overline{D}_2\ (f, W) = \sum_{x \epsilon T \quad UT_2} | W(x, c) - f(x)|^2/(N_1+N_2),$$

the mean square error criterion. Patterson and Womack have called $M(N_1, N_2)$ the mean square error criterion.

Example. Let $T_1 = \{[0, 1]^T, [1, 0]^T, [2, 1]^T\}$ and

$$T_2 = \{[0, 0]^T, [2, 0]^T, [0, -1]^T\}.$$

Then, assuming $W(x, c)$ to be of the form

$$W(x, c) = c_1 x_1^2 + c_2 x_2^2 + c_3 x_1 x_2 + c_4 x_1 + c_5 x_2 + c_6 \qquad (16)$$

and $C(1|2) = C(2|1) = 1$, we can obtain the parameters $c_1, c_2, \ldots, c_6$ by the least square method. Solving the normal equations, we obtain

$$c_1 = -11/8,\ c_2 = 1/4,\ c_3 = 5/8,\ c_4 = 17/8,\ c_5 = 1 \text{ and } c_6 = -1/4$$

and $W(x, c) = (1/8)[2x_2^2 - 11x_1^2 + 5x_1x_2 + 17x_1 + 8x_2 - 2]$.

A decision rule may be defined in order to assign $x \varepsilon \Pi_1$ whenever $W(x, c) > 0$ and to Π_2 otherwise. We may note that such a decision rule correctly classifies all the observations in T_1 and T_2.

In this example, the sets T_1 and T_2 can be separated by a quadric such as parabola, ellipse, hyperbola or circle. That is why a classifier assumed to be of the form (16) of a second degree polynomial in x could classify all the points of T_1 and T_2 correctly. But there are cases when the sets T_1 and T_2 cannot be separated by any such curve. Then a second degree polynomial in x will not be so efficient, that is, the classifier will not be able to classify so many points of T_1 and T_2.

4.2. Potential Function Method

The potential function method was introduced for adaptive classifier design by Aizerman, Braverman and Rozonoer [1]. They assumed that a function $f(x)$, $f(x) > 0$ on T_1 and $f(x) < 0$ on T_2, that is, sign $f(x) = 1$ on T_1 and sign $f(x) = -1$ on T_2, that can be used to separate T_1 and T_2 can be expressed in the form

$$f(x) = c_1 Q_1(x) + \ldots + c_m Q_m(x), \tag{17}$$

where $\{Q_1(x), \ldots, Q_m(x)\}$ is a finite set of orthogonal or orthonormal functions. This function $f(x)$, which can be used as a classifier, can be obtained iteratively using the following algorithm.

$$f_{n+1}(x) = f_n(x) + s(n+1)\, K(x, x(n+1)),\ f_o(x) = 0, \tag{18}$$

where $s(n+1) = \text{sign } f(x(n+1)) - \text{sign } f_n(x(n+1))$, $f_n(x)$ is the approximation to $f(x)$ obtained at the nth step of iteration, $x(n)$ is the nth observation in a sequence of observations of known classification from the

mixture of the populations Π_1 and Π_2 and $K(x, y)$ is a potential function of the form

$$K(x, y) = \sum_{i=1}^{m} Q_i(x)\, Q_i(y). \tag{19}$$

In order to avoid the problem of choice of the set $\{Q_1(x), \ldots, Q_m(x)\}$ Braverman [3] made the following suggestions. Let us assume that $f(x)$ can be represented by

$$f(x) = \sum_{i=1}^{\infty} c_i\, Q_i(x), \quad \sum_{i=1}^{\infty} c_i^2 < \infty, \tag{20}$$

where $\{Q_i(x)\}$ is a complete set of square integrable functions. Then $K(x, y)$ in the above algorithm (18) should be replaced by a function $K(x, y)$ of the form

$$K(x, y) = \sum_{i=1}^{\infty} Q_i(x)\, Q_i(y). \tag{21}$$

Braverman showed that if we take

$$K(x, y) = F\,(d(x, y)), \tag{22}$$

where F is a continuous real function, $d(x, y)$ is any function defining distance between x and y and F has a positive Fourier transform everywhere, then $K(x, y)$ can be represented in the form (21). For example, $K(x, y)$ may be chosen as

$$K(x, y) = \exp\{ - \alpha^2 \sum_{i=1}^{p} | x_i - y_i |^2\}, \tag{23}$$

where α is any finite real number. Braverman has shown that the sequence $f_n(x)$, where x is an observation from Π_1 or Π_2, converges in mean to the function f(x).

Example. In the expression (23) for K(x, y) we choose α=1 and illustrate the construction of $f_n(x)$.

n	x_1	x_2	Member	s(n)	$f_n(x)$
1	0	0	T_2	-1	$-\exp\left(-\Sigma x_i^2\right)$
2	0	1	T_1	2	$2\exp\{-x_1^2-(x_2-1)^2\}+f_1(x)$
3	2	0	T_2	0	$f_2(x)$
4	2	1	T_1	0	$f_2(x)$
5	0	-1	T_2	0	$f_2(x)$
6	1	0	T_1	2	$f_2(x)+2\exp\{-(x_1-1)^2-x_2^2\}$

4.3. Stochastic Approximation Method

Yau and Schumpert [10] have used stochastic approximation method in obtaining an adaptive classifier W(x, c), as in (1), as an approximation to a function f(x) that takes the value 1 on T_1 and -1 on T_2 using G(c) given by

$$G(c) = E\,|\,f(x) - W(X, c)\,|^2,$$

a special case of (5), as a pay-off function. $G(c)$ has a unique minimum, or, the equation $G'(c) = 0$ has a unique solution. But the density $q_1p_1(x) + q_2p_2(x)$ of X from the mixture of Π_1 and Π_2 being unknown, $G(c)$ is unknown. The solution c of $G'(c) = 0$ can be obtained iteratively using the following algorithm.

$$c(n+1) = c(n) - a_n \{c^T(n)\ \phi\ (x(n) - f(x(n))\}\ \Phi\ (x(n)),$$

where $x(n)$ is the nth observation from the mixture of Π_1 and Π_2 $\Phi^T(x) = [Q_1(x), \ldots, Q_m(x)]$ and $\{a_n\}$ is a sequence of real numbers satisfying the conditions

$$\text{(a)}\ \sum_{n=1}^{\infty} a_n = \infty \text{ and } \sum_{n=1}^{\infty} a_n^2 < \infty.$$

$\{1/n\}$ is an example of one such sequence. The convergence of $c(n)$ to c follows from the fact that $c(n)$ forms a Robbins — Monroe process.

References

[1] M. A. Aizerman, E. M. Braverman and L. I. Rozonoer, "The Probability Problem of Pattern Recognition Learning and the Method Potential Functions." Automata and Remote Control, Vol. 25, No. 9, September, 1964.

[2] T. W. Anderson, An Introduction to Multivariate Statistical Analysis. New York; John Wiley and Sons, 1958.

[3] E. M. Braverman, "On the Method of Potential Functions." Automata and Remote Control, Vol. 26, No. 12, 1965.

[4] N. Dunford and J. T. Schwartz, Linear Operators - Part T., New York; Interscience Publishers, Inc., 1967.

[5] Y. Ho and A. Agrawala, "On Pattern Classification Algorithms; Introduction and Survey", IEEE Proc., Vol. 56, No. 12, pp. 2101 - 2124, December, 1968.

[6] J. S. Koford and G. F. Groner, "The Use of an Adaptive Threshold Element to Design a Linear Optimal Pattern Classifier", IEEE Trans. Inform. Theory, Vol. IT - 12, pp. 42 - 52, January, 1965.

[7] N. J. Nilsson, Learning Machines, New York; McGraw Hill, 1965.

[8] J. D. Patterson and B. F. Womack, "An Adaptive Pattern Classification System," IEEE Trans. Syst. Sc. and Cybernatics, Vol. SSC-2, No. 1, pp. 62 - 67, August, 1966.

[9] Y. Z. Tsypkin, Foundations of the Theory of Learning System. New York; Academic Press, 1973.

[10] S. S. Yau and J. M. Schumpert, "Design of Pattern Classifiers With the Updating Property Using Stochastic Approximation." IEEE Trans. Comput. C - 17, No. 9, pp. 861 - 872, September, 1968.

ON RECOGNITION OF WANDERING PATTERNS

J. P. Basu and P. L. Odell

THE UNIVERSITY OF TEXAS AT DALLAS

This research has been supported by Johnson Space Research Center Contract #NAS 9-13512.

ABSTRACT

Two recursive estimators for the current 'mean' of two stochastic processes used in modelling patterns varying over space and/or time have been obtained. A brief survey of models and estimators so far proposed in literature has also been made.

1. Introduction

In the analysis of remotely sensed data, such as multispectral scanner (mss) data, it is usual practice to assume that the data generated come from populations having multivariate normal distribution. When the parameters of the distributions, namely, the means and the covariance matrices, are known, then assuming the misclassification costs to be the same for all populations the Bayes' classifier, the classifier with minimum probability of misclassification, is given by

$$W_{ij}(x) = \log p_i\, x) - \log p_j(x) + \log (q_i/q_j), i\neq j, i, j=1,\ldots,m,$$

where m is the number of populations and $p_i(x)$ is the density and q_i is the prior probability of the ith population. When the parameters are unknown the usual and useful practice in designing classifiers is to replace the unknown parameters in (1) by their respective estimates obtained from a training sample, a sample of known classification. These sample based classifiers, which we shall refer to as estimate plug-in classifiers, do not in general minimize the probability of misclassification. But they are useful in large number of applications and theoretically desirable since we know (Anderson [2]) that as the training sample sizes increase the estimates converge with probability one to the value of the unknown parameter, provided that the observations in a training sample are identically distributed. Thus for large training sample sizes and each training sample having identically distributed observations a sample based classifier may be expected to perform satisfactorily.

The data obtained by remote sensing devices in the earth resources survey come from large areas and over a long period of time. Often the

statistical characteristics of the data have been found to undergo changes over space and time due to variation in spatial and temporal conditions. Thus observations in a training sample collected from a large area and/or over a long period of time are not identically distributed. Therefore, the estimate plug-in classifiers in which the estimates are based on such samples can no longer be expected to perform satisfactorily in classifying patterns that vary over space or time. The performance of these estimate plug-in classifiers can be improved by updating the estimates whenever necessary.

The estimate updating methods are applicable to all populations in the same way. So, without loss of generality, we can discuss these methods in the context of any single population.

The efficiency of the updated estimates depends to a great degree on the accuracy of the statistical model chosen to represent the statistical nature of the varying patterns. But the complexity of the statistical model may on the other hand lead to estimates that are not useful for practical purposes because of computational inconvenience. In this paper we shall consider two general models for estimation and some "quick" methods of estimation based on exponential smoothing techniques (Box and Jenkins [3], Brown [4]).

Throughout the paper we shall denote a sequence of observation vectors from the p- dimensional population Π by $Y_1, Y_2, \ldots, Y_n, \ldots$ and assume

$$Y_k = X_k + W_k, \tag{1}$$

where X_k's are the signals, true patterns or means of Y_k's and W_k's are

p - variate normal random noise vectors distributed independently and identically as N(0, C). Henceforth we shall write L(·) to designate the distribution law of the random vector or vectors in the parenthesis.

2. Autoregressive Model

The pattern X_k can be assumed to be a member of a r-th order autoregressive process. For the sake of simplicity and mathematical tractability we assume that X_k is a member of a first order autoregressive vector process given by

$$X_k - A X_{k-1} = Z_{k-1} \tag{2}$$

where A is a pxp matrix of real numbers, Z_k's are independently and identically distributed as N (0, Σ), Z_k's are independent of W_k's and $L(X_1) = N(0, V)$. The process is known (Box and Jenkins [3]) to be stationary if $|A| < 1$, where $|A|$ denotes the determinant of A, and nonstationary otherwise.

We shall denote by L(P), L(P, Q) and L(P|Q) the probability distribution of P, P and Q jointly and conditional distribution of P given Q respectively. We know that if $L(P, Q) = N(\mu, D)$, $L(P) = N(\mu_1, D_{11})$, $L(Q) = N(\mu_2, D_{22})$ and

$$D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},$$

then $L(P|Q) = N(Q, \phi)$, where

$$Q = \mu_1 + D_{12} D_{22}^{-1} (Q - \mu_2) \tag{3}$$

and $$\phi = D_{11} - D_{12} D_{22}^{-1} D_{21}.$$

Therefore, from (1) and (2) we have

$$Y_1 = X_1+W_1,\ L(Y_1|X_1) = N(X_1, C),\ L(Y_1) = N(0, V+C).$$

and hence $L(Y_1, X_1) = N(0, Q)$, where

$$Q = \begin{bmatrix} V+C & V \\ V & V \end{bmatrix}.$$

From (3) we now have $L(X_1|Y_1) = N(\mu_1, Q_1)$, where

$$\mu_1 = E(X_1|Y_1) = V(V+C)^{-1}Y_1 \tag{4}$$

$$Q_1 = \text{Cov}(X_1, X_1|Y_1) = C(V+C)^{-1}V.$$

Similarly we have

$$Y_2 = X_2+W_2 = AX_1+Z_1+W_2,$$

so, $L(Y_2|Y_1) = N(A\mu_1, AQ_1A^T+\Sigma+C)$ and $L(X_2|Y_1) = N(A\mu_1, AQ_1A^T+\Sigma)$

As in (4), we obtain from (3)

$$\mu_2 = E(X_2|Y_2, Y_1) = C(AQ_1A^T+\Sigma+C)^{-1}A_1 + (AQ_1A^T+\Sigma)(AQ_1A^T+\Sigma+C)^{-1}Y_2$$

and $$Q_2 = \text{Cov}(X_2, X_2|Y_2, Y_1) = C(AQ_1A^T+\Sigma+C)^{-1}(AQ_1A^T+\Sigma)$$

thus $\mu_k = E(X_k | Y_k, \ldots, Y_1) = C(AQ_{k-1}A^T+\Sigma+C)^{-1}A\mu_{k-1}+(AQ_{k-1}$
$+ (AQ_{k-1}A^T+\Sigma)(AQ_{k-1}A^T+\Sigma+C)^{-1}Y_{k-1}$

and $Q_k = \text{Cov}\ (X_k, X_k | Y_k, \ldots, Y_1) = C\ (AQ_{k-1}A^T+\Sigma+C)^{-1}(AQ_{k-1}A^T+\Sigma)$ (5)

From the above model we can obtain the model of Abramson and Braverman [1] by taking $A = aI_p$, where I_p is the pxp identity matrix and a is a real number, and that of Scudder [12] by taking $A = I_p$. If we assume $|A| < 1$, so that the process is stationary, then we obtain the model proposed by Tamura et al [14]. Following Abramson and Braverman, if we assume for slowly varying patterns that $Q_{k-1} = \lambda C$, $\Sigma = \{\lambda^2/(1-\lambda)\}\ C$, $0<\lambda<1$, in order that $Q_{k-1} = Q_k$, then we have from (5)

$$\mu_k = (1-\lambda)\ \mu_{k-1} + \lambda Y_k$$

Writing $\mu_k' = E(X_{k+1} | Y_k, \ldots, Y_1)$, we obtain

$$\mu_k' = E(X_{k+1} | Y_k, \ldots, Y_1)$$

$$= E(X_k+Z_k | Y_k, \ldots, Y_1) = \mu_k.$$

Therefore $\mu_k' = (1-\lambda)\ \mu_{k-1}' + \lambda Y_k$. (6)

The above recursive formula is reminiscent of the exponentially weighted averages to be discussed in section 4.

In the model (2) it is implied that a new pattern is encountered at each observation, in other words, $Y_1, \ldots, Y_n$ are observations on random vectors each having different means. The efficiency of the estimator (5)

will be poor if the changes in pattern are few and far between. In Section 6 we will come back to the question of how to improve the estimates (5).

3. Chernoff - Zacks Model

For analyzing time varying patterns Chernoff and Zacks [6] proposed a model and obtained minimum variance linear unbiased (MVLU) estimate of the current mean. They also obtained Bayes' estimator and an "ad hoc" estimator, a simplified version of Bayes' estimator. But they are unusable for practical purposes, Bayes' estimator due to computational difficulties and the "ad hoc" estimator due to its being based on many restrictive assumptions. Chernoff and Zacks dealt with univariate observations. There we have Y_k, X_i, W_i as univariate random variables. The model assumes

$$Y_i = X_i + W_i, \qquad i = 1, \ldots, n, \tag{7}$$

$$X_i = X_{i+1} + J_i Z_i, \quad i = 1, \ldots, n-1 \tag{8}$$

where W_i's are independently and identically distributed as $N(0, 1)$, J_i's are random variables having the value 1 if there is a change in the mean or pattern X_i between ith and (i+1)th observation and the value 0 otherwise and Z_i is a random variable representing the amount of change when a change takes place.

Let us write

$$\underset{n\times 1}{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \underset{n\times 1}{W} = \begin{bmatrix} W_1 \\ \vdots \\ W_n \end{bmatrix}, \quad \underset{n\times 1}{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_{n-1} \\ 0 \end{bmatrix} \text{ and } \underset{n\times 1}{J} = \begin{bmatrix} J_1 \\ \vdots \\ J_{n-1} \\ 0 \end{bmatrix}. \tag{9}$$

Then noting that we can combine (7) and (8) to obtain the equation

$$Y_i = X_n + W_i + \sum_{k=i}^{n-1} J_k Z_k, \quad 1 \leq i \leq n-1,$$

$$= X_n + W_n, \qquad i=n,$$

we can write

$$Y = X_n e_n + W + MZ, \tag{10}$$

where $M_{nxn} = \begin{bmatrix} J_1 & J_2 & \cdots & J_{n-1} & 0 \\ & J_2 & \cdots & J_{n-1} & 0 \\ & & \cdots & J_{n-1} & 0 \\ 0 & & & & 0 \end{bmatrix}$ and $e_{nx1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$.

If we assume

$$P(J_k=1) = p_k, \quad k=1, \ldots, n-1, \tag{11}$$

and $$L(Z) = N(0, \sigma^2 S)$$

where $$S = \begin{bmatrix} I_{n-1} & 0 \\ 0 & 0 \end{bmatrix}, \tag{12}$$

then we obtain from (10) that

$$L(Y|X_n, J) = N(X_n e_n, Q(J)),$$

where $$Q(J) = I_n + \sigma^2 MM^T. \tag{13}$$

From (11) we have

$$EQ(J) = \sum_{k=1}^{n-1} E[p_k\{Q(J)\,|\,J_k=1\} + (1-p_k)\{Q(J)\,|\,J_k=0\}]$$
$$= I_n + \sum_{k=1}^{n-1} p_k\, E\,(MM^T\,|\,J_k=1)$$
$$= I_n + \sum_{k=1}^{n-1} p_k\, R(n, k), \tag{14}$$

where $R(n, k)$ is a nxn matrix whose upper left submatrix is E_k, a kxk matrix all of whose elements are 1, and the other elements are zero.

When changes in the mean occur almost always or when there are almost no changes, the random vector J can be assumed to be distributed independently of Y. In that case we obtain from (13) and (14)

$$L(Y\,|\,X_n) = N(X_n e_n,\ V_n), \tag{15}$$

where
$$V_n = EQ(J) = I_n + \sum_{k=1}^{n-1} \sigma^2\, p_k\, R(n, k). \tag{16}$$

If μ_n is the current mean, the mean of Y_n, then using standard arguments it can be shown from (15) that

$$\hat{\mu}_n = e_n^T V_n^{-1} Y \,/\, e_n^T V_n^{-1} e_n \tag{17}$$

is a MVLU estimator of μ_n. If $v(n, j)$ denotes the sum of the elements of the jth column of V_n^{-1}, then $\hat{\mu}_n$ can be written in the form

$$\hat{\mu}_n = [\sum_{j=1}^{n-1} v(n, j)Yj + Y_n]/[\sum_{j=1}^{n-1} v(n, j) + 1]. \tag{18}$$

When we have $n+1$ observations, we can obtain in a way similar to (16)

$$V_{n+1} = I_{n+1} + \sum_{k=1}^{n} \sigma^2 p_k \, R(n+1, k)$$

$$= \begin{bmatrix} V_n & 0 \\ 0 & 1 \end{bmatrix} + \sigma^2 p_n \begin{bmatrix} e_n \\ 0 \end{bmatrix} [e_n^T \; 0]$$

We know (Rao [11] p. 29) that if

$$\underset{p\times p}{B} = \underset{p\times p}{A} + \underset{p\times 1}{U} \, \underset{1\times p}{V^T},$$

then $$B^{-1} = A^{-1} - A^{-1}UV^TA^{-1}/(1 + V^TA^{-1} U). \tag{19}$$

Therefore,

$$V_{n+1}^{-1} = \begin{bmatrix} V_n^{-1} & 0 \\ 0 & 1 \end{bmatrix} - \frac{\sigma^2 P_n}{1+\sigma^2 p_n \, e_n^T \, V_n^{-1} \, e_n} K_{n+1},$$

where $$K_{n+1} = \begin{bmatrix} v(n, 1) \\ \vdots \\ v(n, n) \\ 0 \end{bmatrix} [v(n, 1) \; \ldots \; v(n, n) \; 0]. \tag{20}$$

So, $$v(n+1, n+1) = 1$$

$$v(n+1, i) = v(n, i)/[1 + \sigma^2 p_n \sum_{k=1}^{n} v(n, k)], \quad 1 \le i \le n,$$

and $$1 + \sum_{i=1}^{n} v(n+1, i) = [1+(1+\sigma^2 p_n) \sum_{i=1}^{n} v(n, i)]/[1 + \sigma^2 p_n \sum_{k=1}^{n} v(n, k)]$$

Therefore, $$\hat{\mu}_{n+1} = [\sum_{k=1}^{n} v(n+1, k) \, Y_k + Y_{n+1}]/[1 + \sum_{k=1}^{n} v(n+1, k)]$$

$$= \frac{\left[\sum_{i=1}^{n} v(n, i)\right] \hat{\mu}_n + \left[1 + \sigma^2 p_n \sum_{i=1}^{n} v(n, i)\right] Y_{n+1}}{1 + (1 + \sigma^2 p_n) \sum_{i=1}^{n} v(n, i)}$$

$$= (1 - \lambda_n) \hat{\mu}_n + \lambda_n Y_{n+1}, \tag{21}$$

where $$\lambda_n = \{1 + \sigma^2 p_n \sum_{i=1}^{n} v(n, i)\}/\{1 + (1+\sigma^2 p_n) \sum_{i=1}^{n} v(n,i)\}. \tag{22}$$

Using (22) we can now recursively estimate the current mean.

4. Exponentially Weighted Moving Average

In analyzing observations on space and time varying patterns, especially time series data, it has been a popular practice (see Brown [4]) to use exponentially weighted moving average (EWMA) (Box and Jenkins [3]) as the current "mean", location or "level" of the process because of its success in a variety of applications. Let $\{\ldots, Y_{t-1}, Y_t\}$ be observations at .. t-1, t (a stochastic process in discrete time). Then an exponentially weighted moving average (EWMA) is a predictor or forecast of a future level at $t + h$ derived by weighting past observations "exponentially" (or geometrically) and expressed in the form

$$\bar{Y}(t, h; \lambda) = \lambda \sum_{r=0}^{\infty} (1-\lambda)^r Y_{t-r}, \quad |\lambda| < 1 \tag{23}$$

$$= \lambda Y_t + (1-\lambda) \bar{Y}(t-1, h; \lambda), \quad |\lambda| < 1. \tag{24}$$

For h=1, we shall write

$$\bar{Y}(t, 1; \lambda) = \bar{Y}_t(\lambda). \tag{25}$$

Thus from (24) and (25) we have

$$\bar{Y}_t(\lambda) = \lambda Y_t + (1-\lambda)\,\bar{Y}_{t-1}(\lambda) \tag{26}$$

$$= \lambda \sum_{r=0}^{\infty} (1-\lambda)^r\, Y_{t-r}$$

Muth [10] has shown that $\bar{Y}_t(\lambda)$, as given by (26), is optimal for some λ for the following nonstationary moving average process. Let

$$Y_t = X_t + W_t,$$

$$X_t = X_{t-1} + Z_t, \tag{27}$$

where W_t's are independently distributed as $N(0, \sigma^2)$ and Z_t's are independently distributed as $N(0, t^2)$. Then the predictor $\bar{Y}_t(\lambda)$ given by (26) minimizes the error variance

$$E(Y_{t+1} - \bar{Y}_t(\lambda))^2 \tag{28}$$

when λ is given by

$$\lambda = 1 + \frac{t^2}{2\sigma^2} - \frac{t}{\sigma}\left(1 + \frac{t^2}{4\sigma^2}\right)^{1/2}. \tag{29}$$

From (28) it also follows that

$$\bar{Y}_t(\lambda) = E(X_{t+1} \mid Y_t, Y_{t-1}, \ldots). \tag{30}$$

When Y_t is a pxl vector, $L(W_t) = N(0, C)$, $L(Z_t) = N(0, \Sigma)$, then following Muth, it can be shown that $\bar{Y}_t(\lambda)$ minimizes the scatter of error given by

$$E\ (Y_{t+1} - \bar{Y}_t(\lambda)\ (Y_{t+1} - \bar{Y}_t(\lambda))^T$$

provided we take

$$\lambda = 1 + (|\Sigma|/2|C|) - \quad |\Sigma|/|C| \sqrt{1 + |\Sigma|/4|C|}. \tag{31}$$

In passing, it should be noted that the choice of the weight λ is vital in the definition of EWMA. In many situations EWMA's may not be optimal, because the underlying process may not be of the form (27), or even if the process is of the form (27), the matrices Σ and C are unknown, so that it may be impossible to obtain the exact value of λ given by (29) or (31). Even then they are attractive because they are easy to compute.

Motivated by the work of Abramson and Braverman [1], especially by the equation (6) which has been derived from a model of the form (27) and perhaps following the popular practice, Kriegler and Horwitz [8] have proposed that for simultaneously updating the current mean of several populations whose covariance matrices do not change any one of the following three algorithms should be adopted.

(1) Exponentially Weighted Running Estimates

If an observation λ has been recognized as one coming from the class Π_i, then the present mean M_i of this class is to be updated to a new value M_i', where

$$M_i' = (1-\lambda)\ M_i + \lambda Y, \tag{32}$$

and λ is a constant.

Kriegler and Horwitz have recommended that λ should be given a value between 0.001 and 0.003. Evidently such small λ will heavily weigh the past observations, especially the present mean M_i and will attach very little weight to the current observation. Unless the patterns are extremely slowly varying, such small λ will make the updated mean M_i' a very poor representative of the current pattern.

(2) Exponentially Weighted Running Estimates With Interaction

Assuming that at any time the mean M_i of the class Π_i is related to the sum M of means of all M classes by the equation

$$M = mp_i M_i, \tag{33}$$

where p_i is constant for all time, the present mean M_j of any clsss Π_j should be updated to the value M_j', where

$$M_j' = p_i M_i'/p_j \tag{34}$$

and M_i' is the current updated mean of Π_i, updated according to (32).

The assumption (33) is hard to justify.

(3) Posterior Probability Weighted Estimates

Let K_i denote the fixed covariance matrix of Π_i. Then according to this algorithm an observation Y_t is assigned to the class Π_i if for some preassigned threshold value t_1,

$$L_i(Y_t) > L_j(Y_t) \text{ and } Q_i(Y_t) < t_1,$$

where $$L_i(Y_t) = |K_i|^{-\frac{1}{2}} \exp\{-Q_i(Y_t)/2\}$$

and $\quad Q_i(Y) = (Y_t - M_i)^T K_i^{-1} (Y_t - M_i).$

The mean M_i should be updated to M_i' given by

$$M_i' = 1 + \gamma R_i(Y_t)(Y_t - M_i) \tag{35}$$

$$= \{1 - \lambda(t, i)\} M_i + \lambda(t, i)\, Y_t,$$

where $R_i(Y_t) = L_i(Y_t) / \sum_{j=1}^{m} L_j(Y_t)$

and $\lambda(t, i) = \gamma R_i(Y_t)$, γ being a constant $(0<\gamma<1)$.

The updating formula (21) based on Chernoff - Zacks model and formula (35), both resemble EWMA, except that the weights are now variable and to be updated each time. The predictor (21) has been shown to be MVLU. But the predictor M_i' given by (35) may not enjoy any such property. It is not known for what kind of process M_i' will be optimal in some sense. Kriegler and Horwitz have suggested some values of t_1, t_2 and γ on empirical grounds, but have failed to give a specific rule for the selection of these quantities. The performances of the predictors (35) in different situations are yet to be determined and compared.

The work of Kriegler and Horwitz [8] motivated Chang [5] to propose updating formulae for the mean and covariance matrix using variable weights. Let $Y_{j1}, \ldots, Y_{jn(j)}$ denote the observation from the jth training field, a set of identically distributed random vectors. Then the updated mean M_j' is given by

$$M_j' = \gamma_j M_{j-1}' + (1-\gamma_j) M_j, \tag{36}$$

where $M_j = (Y_{j1} + \ldots + Y_{jn(j)})/\, n(j)$,

$\gamma_j = a_{j-1}\, N_{j-1}/\, N_j$,

$N_k = n(1) + \ldots + n(k)$

and a_{j-1} is a number satisfying $0 \leq a_{j-1} \leq N_j/N_{j-1}$ and to be guessed from the nature of the data. The covariance matrix is to be updated to K_j where

$$K_j = \{N_j/(N_j-1)\}\ (Q_j' - M_j' M_j'^{T}), \tag{37}$$

$$Q_j' = \gamma_{j-1}\, Q_{j-1}' + (1-\gamma_{j-1})\, Q_j,$$

and $$Q_j = \sum_{i=1}^{n(j)} Y_{ji}\, Y_{ji}^{T}/\, n(j).$$

When the patterns are varying, the updated "mean" M_j' given by (36) is not an unbiased estimate of the current mean. It is not known in what way these updated estimates are optimal. The other disadvantage of this algorithm is that it leaves the choice of γ_i to the guess of the user.

5. Updating By Kalman Filters

Consider the dynamic relationship

$$Y_k = H_k S_k + W_k \tag{38}$$

$$S_k = AS_{k-1} + Z_{k-1}$$

satisfied by the pxl observation vectors $Y_1, \ldots, Y_k, \ldots$ and the qxl state vectors $S_1, \ldots, S_k, \ldots$, where W_k's are pxl random

noise vectors independently distributed as $N(0, C)$, Z_k's are qx1 random vectors distributed independently of each other and of W_k's as $N(0, Q)$ and the pxq matrix H_k and qxq matrix A may be time varying, but are known matrices of real numbers. It is well known (Lee [9]) that an estimate $\hat{S}_k$ of S_k given $Y_1, \ldots, Y_k$ such that

$$E(\hat{S}_k) = E(S_k \mid Y_k, \ldots, Y_1) \tag{39}$$

and $E[(\hat{S}_k - S_k)^T (S_k - S_k)]$ is minimum

can be obtained, using discrete Kalman filter, recursively as

$$\hat{S}_k = A\hat{S}_{k-1} + P_k H_k^T C^{-1}[Y_k - H_k A\hat{S}_{k-1}], \tag{40}$$

$$P_k = [(A P_{k-1} A^T + Q)^{-1} + H_k^T C^{-1} H_k]^{-1}. \tag{41}$$

$$= \text{Cov}\,(S_k, S_k \mid Y_k, \ldots, Y_1)$$

Assuming covariance matrices of all populations to be the same and not undergoing any change, Crane [7] used the estimation method by Kalman filters for simultaneously updating the means of the populations $\Pi_1, \ldots, \Pi_m$. Crane considered the following model:

$$Y_k = H_k S_k + W_k, \tag{42}$$

$$S_k = S_{k-1} + Z_{k-1}, \tag{43}$$

$$H_k = M_k \otimes I_p, \tag{44}$$

where M_k is a mx1 matrix with ith row (i=1, ..., m) as 1 if Y_k has been recognized to be coming from Π_i and zero otherwise and $A \otimes B$ denoting the Kronecker product of the matrices A and B (Anderson [2]). It has also been assumed that

$$Q = R \otimes C, \tag{45}$$

where $$\underset{mxm}{R} = r_1 \begin{bmatrix} 1 & r_2 & \cdots & r_2 \\ r_2 & 1 & \cdots & r_2 \\ \vdots & & \ddots & \vdots \\ r_2 & & & 1 \end{bmatrix}. \tag{46}$$

The state vector S_k is a m px1 vector of m stacked subvectors of px1 vectors of present means (prior to updating). Let $X_k^{(i)}$ and $Z_k^{(i)}$ denote the ith px1 vector component of S_k and Z_k respectively. Then assuming Y_k to be an observation from Π_i, we obtain from (42), (43) and (44) that

$$Y_k = X_k^{(i)} + W_k \tag{47}$$

$$X_k^{(i)} + X_{k-1}^{(i)} + Z_{k-1}^{(i)} . \tag{48}$$

This is the process for which EWMA s are optimal. Thus it is evident that Crane's model is a generalization of the above model so that in simultaneously updating the means of all populations their interaction has been taken into account.

Noting that

$$[(P_{k-1} + Q)^{-1} + H_k^T C^{-1} H_k]^{-1} H_k^T C^{-1}$$

$$= (P_{k-1}+ Q) H_k^T [H_k (P_{k-1}+ Q) H_k^T + C]^{-1},$$

we now obtain from (40) that

$$\hat{S}_k = \hat{S}_{k-1} + (P_{k-1}+ Q) H_k^T [H_k (P_{k-1}+ Q) H_k^T + C]^{-1}(Y_k - H_k \hat{S}_{k-1}). \quad (49)$$

Further assuming $L(S_1) = N(0, P_o)$ and

$$\underset{mpxmp}{P_o} = \underset{mxm}{T_o} \otimes \underset{pxp}{C}, \quad (50)$$

we can obtain from (45) that

$$P_1+Q = (P_o^{-1} + H_1^T C^{-1} H_1)^{-1} + Q$$

$$= (T_o^{-1} + M_1 M_1^T)^{-1} \otimes C + R \otimes C$$

$$= \{T_o - \frac{T_o M_1 M_1^T T_o}{1+M_1^T T_o M_1}\} \otimes C + R \otimes C$$

$$= T_1 \otimes C$$

and $$P_2+Q = [(P_1+Q)^{-1} + H_2^T C^{-1} H_2]^{-1} + R \otimes C$$

$$= (T_1^{-1} + M_2 M_2^T)^{-1} \otimes C + R \otimes C$$

$$= \{T_1 - \frac{T_1 M_2 M_2^T T_1}{1+M_2^T T_1 M_2}\} \otimes C + R \otimes C$$

$$= T_2 \otimes C,$$

where $$T_1 = T_o - \frac{T_o M_1 M_1^T T_o}{1+M_1^T T_o M_1} + R$$

and $$T_2 = T_1 - \frac{T_1 M_2 M_2^T T_1}{1+M_2^T T_1 M_2} + R .$$

By induction, we can obtain

$$\underset{mpxmp}{P_k} + \underset{mpxmp}{Q} = \underset{mxm}{T_k} \otimes \underset{pxp}{C},$$

where $$T_k = T_{k-1} - \frac{T_{k-1} M_k M_k^T T_{k-1}}{1 + M_k^T T_{k-1} M_k} + R \qquad (51)$$

Using (51) we can simplify (49) and obtain

$$\hat{S}_k = \hat{S}_{k-1} + [(1+M_k^T T_{k-1} M_k)^{-1} T_{k-1} M_k \otimes I_p] (Y_k - H_k \hat{S}_{k-1}). \qquad (52)$$

Crane has suggested some "ad hoc" value for r_1 and r_2 in (46) in order to define the interaction matrix R. But it seems more reasonable to estimate them from a part of the sample.

The above estimation procedure will be clear from the following numerical example.

Example. Let $m = 2$, $p = 2$, $T_o = I_2$, $C = I_2$, $r_1 = 0.01$, $r_2 = 0.02$, $Y_1^T = [1\ 2] \in \Pi_1$, $Y_2^T = [-1\ 1] \in \Pi_2$. Then

$$M_1^T = [1\ 0],\ m_2^T = [0\ 1],$$

$$M_1^T T_o M_1 = 1,$$

$$T_1 = I_2 - \{I_2 \begin{bmatrix}1\\0\end{bmatrix} [1\ 0]\ I_2\}/\ 2 + 0.01 \begin{bmatrix}1 & 0.02\\0.02 & 1\end{bmatrix}$$

$$= \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} - \begin{bmatrix}0.5 & 0\\0 & 0\end{bmatrix} + \begin{bmatrix}0.01 & 0.002\\0.002 & 0.01\end{bmatrix}$$

$$= \begin{bmatrix}0.51 & 0.002\\0.002 & 1.01\end{bmatrix};$$

$$M^T{}_2 T_1 M_2 = 1.01,$$

$$T_2 = T_1 - \{T_1 \begin{bmatrix}0\\1\end{bmatrix} [0\ 1]\ T_1\}/\ 2.01 + R$$

$$= T_1 - \frac{1}{2.01} \begin{bmatrix}0 & 0.002\\0 & 1.01\end{bmatrix} \begin{bmatrix}0.51 & 0.002\\0.002 & 1.01\end{bmatrix} + R.$$

6. Detection of Change in Mean

The models on which we have based our estimation process so far implicitly assume that the means change almost at each observation. If this is the case, then the updating of means using the procedures so far discussed is just. But when we are considering spatial variation and

have observations from points spaced not too far from each other, then it is more likely that the sequence of observations $Y_1, Y_2, \ldots, Y_n$ can be grouped as $\{Y_1, Y_2, \ldots, Y_{k_1}\}, \{Y_{k_1+1}, Y_{k_1+2}, \ldots, Y_{k_2}\}, \ldots$ such that members in the same group have the same mean while members in different groups have different means. In that case it will be more reasonable to use the arithmetic mean $(Y_{k_i+1} + \ldots + Y_{k_{i+1}})/(k_{i+1} - k_i)$ as the mean of $Y_j (k_i+1 \leq j \leq k_{i+1})$ than to use the mean obtained by the updating process. But often the points $K_1, k_2, \ldots$ at which the changes in mean take place is unknown. Therefore we have to use some statistical test in order to determine the points of change.

When $Y_1, Y_2, \ldots, Y_n$ are pxl random vectors independently distributed normally with the same covariance matrix, say I_p, then for testing the hypothesis

$$H_o : \mu_1 = \mu_2 = \ldots = \mu_n \quad \text{(equality of means)}$$

against $H_1 : \mu_1 = \ldots = \mu_r \neq \mu_{r+1} = \ldots = \mu_n$,

where the change point r is unknown and $1 \leq r \leq N-1$, Sen and Srivastava [13] have used the test statistic U^* given by

$$U^* = N^{-2} \sum_{i=1}^{n-1} \left(\sum_{j=i}^{n-1} (Y_{j+1} - \bar{Y}) \right)^T \left(\sum_{j=i}^{n-1} (Y_{j+1} - \bar{Y}) \right),$$

where $\bar{Y} = (Y_1 + \ldots + Y_n)/N$. They have also given the asymptotic percentile of U^* for $p=2$ through $p=8$. Chernoff and Zacks [6] have also given a test for $p=1$. For other references of work in this connection we refer to [13].

References

[1] N. Abramson and D. Braverman, "Learning to recognize patterns in a random environment," IRE Trans. Inf. Theory, Vol. IT-8, pp. 58-63, September, 1962.

[2] T. W. Anderson, Introduction to Multivariate Statistical Analysis, New York, Wiley, 1958.

[3] G. E. P. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control, San Francisco, Holden - Day, 1970.

[4] R. G. Brown, Smoothing, Forecasting and Prediction, New Jersey, Prentice-Hall, 1963.

[5] C. Y. Chang, "Weighted adaptive algorithm for estimation of Gaussian distribution parameters," to appear in IEEE Trans. SSC.

[6] H. Chernoff and S. Zacks, "Estimating the current mean of a normal distribution which is subjected to a change in time," Ann. Math. Statist., Vol. 25, pp. 999-1018, 1964.

[7] R. B. Crane, "Adaptive processing with decision-directed Kalman filter and feature extraction of multispectral data," Technical Report, NASA CR-ERIM 190100-31-T, ERIM, P. O. Box 618, Ann Arbor, Michigan, 1973.

[8] F. J. Kriegler and H. M. Horwitz, "Investigations in adaptive processing of multispectral data," Technical Report, NASA CR-ERIM 31650 - 151 - T, ERIM, Michigan, 1973.

[9] R. C. K. Lee, Optimal Estimation, Identification and Control, Cambridge, Mass., MIT Press, 1964.

[10] J. F. Muth, "Optimal properties of exponentially weighted forecasts," Journal of Amer. Stat. Asso., Vol. 55, pp. 299-306, June, 1960.

[11] C. R. Rao, Linear Statistical Inference and Its Applications, New York, Wiley, 1965.

[12] H. J. Scudder, "Probability of error of some adaptive pattern recognition machines," IEEE Trans. Inf. Theory, Vol. IT-11, pp. 363-371, July, 1965.

[13] A. K. Sen and M. S. Srivastava, "On multivariate tests for detecting change in mean," Sankhya, Ser. A, Vol. 35, pp. 173-186, 1973.

[14] S. Tamura, S. Higuchi and K. Tanaka, "On the recognition of time varying patterns using learning procedures," IEEE Trans. Inf. Theory, Vol. IT-17, pp. 445-452, July, 1971.

Addendum to Papers 1, 4 and 5

Consider the matrix P for the two crop situation. Say,

$\hat{P} = \begin{bmatrix} \hat{P}(1|1) & \hat{P}(1|2) \\ \hat{P}(2|1) & \hat{P}(2|2) \end{bmatrix}$, where $\hat{P}(1|1) = x/N_1$ $\hat{P}(2|2) = y/N_2$ and $\hat{P}(2|1) = 1 - \hat{P}(1|1)$, $\hat{P}(1|2) = 1 - \hat{P}(2|2)$. x is the number correctly classified into population one and y is the number correctly classified into population two. Since X and Y are independently distributed as binomial variates, the probability that $\hat{P}$ is singular is positive. This is illustrated for the case $N_1=N_2 = 3$ where p_i represents the probability of correctly classifying an observation in population i.

Since $\hat{P}$ is singular iff $xy - (3-x)(3-y) = 0$, iff $x+y = 3$, there are 4 points yielding $\hat{P}$ singular. They are (1,2), (2,1), (0,3), (3,0) and the probability $\hat{P}$ is singular is $9\, p_1q_1^2p_2^2q_2 + 9\, p_1^2q_1p_2q_2^2 + q_1^3p_2^3 + p_1^3q_2^3$. In general, if $N_1=N_2$ there are N_2+1 points which yield $\hat{P}$ singular.

Hence, an alternate method must be used to estimate $\hat{P}$. One approach is to estimate μ_i, Σ_i for each population from training samples and then estimate P(i|j) by $\hat{P}(i|j) = \int_{R_i} p_j(x;\hat{\mu}_j,\hat{\Sigma}_j)\, dx$, where R_i is the region corresponding to an observation being classified into the ith population, and p_j is the (continuous) density function of the jth population. The probability that $\hat{P}$ is non-singular with probability one can be established using advanced probabilistic arguments.

Another approach is to use the pseudoinverse of $\hat{P}$ rather than the inverse. However, this approach would affect all the analyses done and the results may be difficult to interpret. This approach is probably the best but needs to be studied to assure that subtleties have not been ignored.

Finally, one could make the analysis conditional on the fact that $\hat{P}$ is non-singular. Formulas for variances and mean squared error of $\hat{\hat{P}}$ would necessarily be modified to incorporate this condition; yet the analysis would conceptually be straightforward and follow the arguments now presented.