PURDUE UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING
ELECTRONIC SYSTEMS RESEARCH LABORATORY

LEARNING PROBABILITY SPACES FOR CLASSIFICATION AND RECOGNITION OF PATTERNS WITH OR WITHOUT SUPERVISION

by

J. C. Hancock and E. A. Patrick

Technical Report No. TR-EE 65-21
Sponsored by National Aeronautics and Space Administration
under Contract NsG-553

NOVEMBER 1965
LAFAYETTE, INDIANA

RESEARCH GRANT

No. NsG-553

PRF 3823

LEARNING PROBABILITY SPACES FOR CLASSIFICATION AND RECOGNITION OF PATTERNS WITH OR WITHOUT SUPERVISION

for

NATIONAL AERONAUTICS AND SPACE

ADMINISTRIATION

WASHINGTON 25, D. C.

by

J.C. Hancock, Principal Investigator

E.A. Patrick

School of Electrical Engineering

Purdue University

Lafayette, Indiana

ACKNOWLEDGMENTS

I wish to thank Professor John C. Hancock for valuable guidance and continued leadership during the course of this research. Thanks are also due Professors J. M. McFadden, King-Sun Fu, and William J. Studden for their helpful suggestions, and to students who assisted with computer programming.

The financial sponsorship for this research by N.A.S.A. Grant (NsG 553) is gratefully acknowledged.

This research is dedicated to my father, A. L. Patrick.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

Symbol	Description
X_s	sth, ℓ-dimensional vector sample
$\{X_s\}_{n-v+1}^{n}$	sequence of v samples, $X_{n-v+1}, X_{n-v+2}, \ldots, X_n$, taken at v different observations
$X_s = \{X_{s_k}\}_1^v$	sequence of v samples taken at the sth observation
$F(X)$	cumulative distribution function (c.d.f.) for X
$F(X\|B)$	parameter-conditional c.d.f.
$F(X\|\omega_i)$	ith class-conditional c.d.f.
$F(X\|\omega_i, B_i)$	ith class, parameter-conditional c.d.f.
ω_i	ith pattern class
Ω	parameter space
M	number of pattern classes
ω_i^s	event ith class was active on sth sample
$\{P_i\}_1^M$	M pattern class probabilities
$\{P(\pi_r)\}_1^W$	W partition probabilities
B_i	vector characterizing ith class-conditional c.d.f.
B_{M+1}	$B_{M+1} = \{P_i\}_1^M$, if W = M
B	$B = B_1 \cup B_2 \cup \ldots \cup B_M \cup B_{M+1}$, if W = M
$F(B)$	a priori c.d.f. for B
B_{i_o}, B_o	true vectors
more generally,	
B_{W+1}	$\{P(\pi_r)\}_1^W$

Symbol	Description
B	$B = B_1 \cup B_2 \cup \ldots \cup B_W \cup B_{W+1}$
B_r	vector characterizing rth partition-conditional c.d.f.

Fixed Bin Model Notation

R	number of quantizing levels for each dimension of X
B_ξ	bin ξ, one of the ℓ-dimensional quantum levels
p_ξ^o	amount of probability from $F(X)$ in B_ξ.
p_ξ^i	amount of probability from $F(X\|\omega_i)$ in B_ξ.
$\underline{P}^o$	$= (p_1^o, p_2^o, \ldots, p_{R^\ell}^o)$
$\underline{P}^i$	$= (p_1^i, p_2^i, \ldots, p_{R^\ell}^i)$
$\mathcal{F}_P$	family of multinomial c.d.f.'s
$\mathcal{F}_{TP}$	family of multinomial c.d.f.'s differing only by translational vectors
$\mathcal{F}_{STP}$	family of symmetrical multinomial c.d.f.'s differing only by translational vectors
V_s	vector of relative frequence in the R^ℓ bins due to sample X_s; $V_s = (v_{s_1}, v_{s_2}, \ldots, v_{s_{R^\ell}})$
s	indicates the sth observation
k	the kth member of a particular observation

ABSTRACT

This thesis is concerned with nonsupervisory problems which arise in the design of numerous types of detection systems. A rather general approach is given which differs from approaches taken by other investigators in that (a) the solution is formulated to include nonparametric as well as parametric knowledge, (b) the definition of the nonsupervisory problem is extended to a class of nonsupervisory problems, and (c) it is recognized that a certain minimum amount of a priori knowledge is required for a solution to exist.

The approach begins by showing that when samples are not classified, the probability distribution of the samples is a mixture c.d.f. A mixture c.d.f. is constructed by utilizing the a priori knowledge available. It is then possible to determine if a sufficient amount of a priori knowledge is available for a solution to exist. By solution is meant that a system exists minimizing sample-conditional probability of error (or, more generally, sample-conditional risk) and converging to the minimum probability of error system.

Histogram and empirical c.d.f. concepts are defined for nonsupervisory problems. Furthermore, it is shown that classical results for Bayes estimates, maximum likelihood estimates, etc. can be applied to nonsupervisory problems.

Computer simulated results verifying the approach are given for several examples.

CHAPTER I

INTRODUCTION

1.1 The Problem

This thesis is concerned with the nonsupervisory problem (i.e. adapting without a teacher) which arises in the design of numerous types of detection systems. Given here is a description of a rather general approach which differs from approaches taken by other investigators in that (a) the solution is formulated to include nonparametric as well as parametric knowledge, (b) the definition of the nonsupervisory problem is extended to a class of nonsupervisory problems, and (c) it is recognized that a certain minimum amount of a priori knowledge is required for a solution to exist.

The study begins with a treatment of how a priori knowledge is taken into account when processing a sequence of vector samples. This a priori knowledge could include knowledge of cumulative distribution functions, possible families of cumulative distribution functions, the number of pattern sources, and any constraints on parameters. The system objective is formulated in such a way that it can be an optimum one which minimizes conditional risk or conditional probability of error or one of a variety of suboptimum applications.

If, as the number of observations becomes large, the system is to converge in the limit to the system obtained when all statistics are known, a ceratin minimum amount of a priori knowledge is required. This minimum amount of a priori knowledge must guarantee that the system will converge; this is equivalent to saying that the parameters characterizing the cumulative distribution function of the observations must be identifiable. If these parameters are identifiable, it is then possible to show that a priori probability laws defined on fixed but unknown parameters are not required.

It is assumed that a sequence of ℓ dimensional samples are presented to a receiver as denoted by

$$X_1,\ X_2,\ldots,X_n:\quad \{X_s\}_1^n \tag{1.1}$$

where X_s is a representative ℓ dimensional vector sample:

$$X_s = (x_{s_1},\ x_{s_2},\ldots,x_{s_\ell}) \tag{1.2}$$

We assume that a cumulative distribution function (c.d.f.) $F(X_s)$ exists. If the form of this distribution function is specified by a vector set of parameters B, then we write this functional form $F(X_s|B)$ and call it the parameter-conditional distribution function. Further, if ω_i is the pattern class or source acting to produce X_s where there are M possible pattern classes, this is denoted as the event ω_i^s. The c.d.f. of X_s, given ω_i^s and a vector B_i, is $F(X_s|\omega_i^s,B_i)$. For convenience we drop the superscript s when it will not cause confusion, and write $F(X_s|\omega_i,B_i)$, meaning it is given that the ith pattern class is acting to produce X_s. $F(X_s|\omega_i,B_i)$ will be called the ith class, parameter-conditional c.d.f. In the nonsupervisory problems considered in this thesis, the family $\{F(X_s|\omega_i,B_i)\}$ will be assumed known a priori.

More generally, let X be any sequence of v samples of the n samples-for example $X = \{X_s\}_{n-v+1}^n$. Let W be the number of possible ways that M classes could be active to cause the v samples. Call the rth way the rth partition, π_r. Then $\{F(X|\pi_r,B_r)\}$ is the family of rth partition, parameter-conditional c.d.f.'s. The definition of this latter family allows the extension of the nonsupervisory problem to a class of nonsupervisory problems. In the literature survey which follows we will be concerned only with the former family, i.e., $\{F(X_s|\omega_i,B_i)\}$.

Define the probability of the event that the ith class is active on the sth sample by $P(\omega_i^s)$. If the probability of this event is independent of the sample number, then $P(\omega_i^s) = P_i$, an assumption made throughout this thesis. The set $\{P_i\}_1^M$ is called the set of mixing parameters corresponding to M members of the family $\{F(X_s|\omega_i,B_i)\}$. More generally, when there are W partitions, $\{P(\pi_r)\}_1^W$ is the set of mixing parameters corresponding to W members of the family $\{F(X|\pi_r,B_r)\}$. If X is indexed, corresponding to a set of sequences of v samples-for example $X_1 = \{X_{1_k}\}_1^v$, $X_2 = \{X_{2_k}\}_1^v, \ldots, X_n = \{X_{n_k}\}_1^v$ we assume that the mixing parameters for each sequence are $\{P(\pi_r)\}_1^W$, independent of the sequence number.

If the samples $\{X_s\}_1^n$ are statistically independent given their c.d.f., then the samples are parameter-conditionally independent and we have

$$F(X_n|B,\{X_s\}_1^{n-1}) = F(X_n|B), \text{ for all n.} \tag{1.3}$$

1.2 Literature Survey

An optimality criterion frequently used is as follows: Given a sequence of ℓ dimensional samples $\{X_s\}_1^{n-1}$, make a decision as to which of M classes is active to cause sample X_n. This decision is made by a decision function obtained with a system constraint of minimum sample-conditional risk. The word sample is used here to make clear that we are talking about risk conditioned on the past samples, $\{X_s\}_1^{n-1}$.

It is desirable that this sample-conditional risk or sample-conditional probability of error become <u>stable</u> as n becomes large. Even more desirable is that the stable point be identical to that obtained if all the vector parameters in B were known. That is, it is desirable that the performance of an adaptive system <u>converge</u> uniquely to that of a system minimizing risk or probability of error. We therefore make a distinction between a stable system and a stable system which converges, the latter implying convergence to the unique system obtainable had all the parameters characterizing the system been known.

A suboptimum system is defined as a system which minimizes probability of error when $n \to \infty$, but which has a sample-conditional probability of error greater than optimum. A suboptimum system possibly could be better than an optimum system when the system complexity, cost, etc. are taken into account.

Abramson and Braverman[1] considered an example where it is known which class is active to cause sample X_s, $s = 1,2,\ldots,n_i$ (i.e., the samples are supervised). That is, the a priori knowledge includes knowledge that

$$F(X_s|B) = F(X_s|\omega_i, B_i), \text{ i known, } s = 1,2,\ldots,n_i \tag{1.4}$$

Further, it is known that the family $\{F(X_s|\omega_i,B_i)\}$ is a multidimensional gaussian family, with only the mean vector m_i (in B_i) unknown for each member. If M groups of supervised samples are taken corresponding to M pattern classes and if all samples are parameter-conditionally independent, then

$$f(\{X_s\}_1^n|B) = \prod_{i=1}^{M} \prod_{s=n_{i-1}}^{n_i} f(X_s|\omega_i,B_i) \tag{1.5}$$

where $n = n_1 + n_2 + \ldots + n_M$. Since the a priori knowledge includes knowledge of the family and of M, Eq. (1.5) is a known function of B. In this example they also assumed that the a priori knowledge includes a c.d.f. F(B). Using this a priori knowledge, they obtained a system minimizing the sample-conditional probability of error.

Keehn[15] extended the work of Abramson and Braverman to the case where the family is multivariate guassian and where the mean vector and covariance matrix are unknown. He carefully defined c.d.f.'s $F(B_i)$ for all i such that the a posteriori c.d.f. of B_i, for each i, is reproducing[20].

Daly[3] investigated a nonsupervisory system where the classification of the samples is unknown. A priori knowledge includes: knowledge that there are M classes with a single class active causing each sample, that the family $\{F(X_s|\omega_i,B_i)\}$ is known, the set of mixing parameters $\{P_i\}_1^M$ are known, and a c.d.f. F(B) is available. Daly computed the sample-conditional risk using this a priori knowledge obtaining, in particular, the decision function which minimizes the sample-conditional probability of error for decision on sample X_n. This decision function computes the sample-conditional density functions, $f(X_n,\omega_i|\{X_s\}_1^{n-1})$, $i = 1,2,\ldots,M$. His computation for $f(X_n, \omega_i|\{X_s\}_1^{n-1})$ is a sum of $M^{(n-1)}$ terms, thus requiring rapidly increasing computer memory. Daly indicated that the system is stable as n becomes large; however, he did not show convergence. In general this solution does not converge, and additional a priori constraints are required to assure convergence. The approach described in this thesis provides for using these additional constraints.

Fralick[2,14], looking for an iterative solution to Daly's problem, obtained an iterative form assuming that if B_i characterizes $F(X_s|\omega_i,B_i)$ and B_j characterizes $F(X_s|\omega_j,B_j)$, then $F(B_i|\{X_s\}_1^{n-1},B_j) = F(B_i|\{X_s\}_1^{n-1})$. Fralick's result is in general suboptimum since, in general, $F(B_i|\{X_s\}_1^{n-1}, B_j) \neq F(B_i|\{X_s\}_1^{n-1})$, $j \neq i$. This condition is true when B_j is known and M = 2, which, with $B_i = B_1$ and $B_j = B_2$, corresponds to the binary on-off case without supervision.

Hancock and Patrick[19,16] showed that the desired a posteriori probability density $f(B_i|\{X_s\}_1^{n-1})$ is either of the growing form, or equivalently is computed by integrating the joint density $f(B|\{X_s\}_1^{n-1})$ with respect to all vectors except B_i, where $f(B|\{X_s\}_1^{n-1})$ has an iterative form. Their result is that

$$f(B_i|\{X_s\}_1^{n-1}) = \int \ldots \int \prod_{j \neq i} d\,B_j \frac{\left[\sum_{j \neq i} P_j\, f(X_{n-1}|\omega_j,B_j) + P_i f(X_{n-1}|\omega_i B_i)\right]}{f(X_{n-1}|\{X_s\}_1^{n-2})} f(B|\{X_s\}_1^{n-2}) \quad (1.6)$$

Equation (1.6) is the Bayes solution for a "mixture" of M class, parameter-conditional c.d.f.'s. This basic result, obtained by Hancock and Patrick, includes Fralick's result as a special case. Equation (1.6) is the result for the specific mixture considered, one of a class of mixtures considered herein, and is an introduction to the parameter-conditional mixture approach to nonsupervisory problems considered in this thesis.

Cooper and Cooper[4] considered the binary ($M = 2$) case with the family $\{F(X_s|\omega_i,B_i)\}$ one dimensional gaussian. They obtained moment estimators for $B_{i_o} = (m_{i_o}, P_{i_o})$ with $\sigma_{1_o} = \sigma_{2_o}$ known, and maximum likelihood estimators for $B_{i_o} = m_{i_o}$ with $P_{i_o} = 1/2$ known, $i = 1,2$. Patrick and Hancock[7], using a different approach obtained maximum likelihood estimators for the more general case where $B_{i_o} = (m_{i_o}, \sigma_{i_o}, P_{i_o})$, all entries in B_{i_o} being unknown.

Some of the first work on applying a histogram, approximating a class-conditional c.d.f. $F(X_s|\omega_i)$, to adaptive communication systems was done by Sebestyen[5,8]. He considered only supervised samples with a single class active on each sample.

Patrick and Hancock[7,6] applied a histogram, approximating a class-conditional c.d.f. $F(X_s|\omega_i)$, to the nonsupervisory problem. They presented computer simulated results[7] for the rate of convergence of a binary system where the a priori knowledge includes knowledge that $F(X_s|\omega_i)$ is symmetrical, $P_{i_o} = 1/2$ and is known, and, that there is an appropriately large signal-to-noise ratio. They compared this rate of convergence with that obtained by two other approaches using moment estimators and maximum likelihood estimators, respectively.

Robbins[12] considered estimators for P_{i_o} with $F(X_s|\omega_i,B_i)$ known, $i = 1,2,\ldots,M$. His estimators are approximations to maximum likelihood estimators, obtained in Chapter V of this thesis, when $F(X_s|\omega_i,B_i)$ is gaussian, and perform badly.

Teicher[9,10] defined a mixture and identifiability and gave a theorem giving sufficient conditions for a mixture to be identifiable. In appendix A we include and give a simple extension of Teicher's work, and define a parameter-conditional mixture which is a useful concept for applying Bayes Theorem to mixtures. In addition we state a theorem and several propositions giving sufficient conditions for a parameter-conditional mixture to be identifiable. One of Teicher's propositions[10], for example, states that a finite mixture of one-dimensional gaussian c.d.f.'s is identifiable if the class-conditional c.d.f.'s can be ordered such that $\sigma_i > \sigma_j$, $i < j$, or if $\sigma_i = \sigma_j$, $m_i < m_j$. An extension of Teicher's proposition in Appendix A gives sufficient conditions for the multidimensional gaussian case.

The work by Daly and Fralick, discussed previously, does not consider identifiability or system constraints assuring a unique solution. The parameter-conditional mixture approach, considered in this thesis, does provide for utilizing such constraints.

1.3 Approach and Contributions

In this thesis the approach to the nonsupervisory problem begins by showing that, when samples are not classified, the probability distribution of the samples is a mixture c.d.f.[9,10] A mixture c.d.f. is constructed by utilizing the a priori knowledge available. If, for example, the a priori knowledge included the classification of the samples, then the c.d.f. of the samples would be a degenerate mixture c.d.f. as in (1.5). In this sense, classification of the samples is a priori knowledge used in constructing the c.d.f. of the samples.

The overall contribution of this mixture approach to nonsupervisory problems is that sufficient amounts of a priori knowledge for a solution to exist can be determined. As an example, for the nonsupervisory problem considered by Daly we show that if the family $\{F(X_s|\omega_i,B_i)\}$ is one dimensional gaussian with $B_i = (m_i, \sigma_i)$, it is sufficient that all the means be unequal in order for a solution to exist. In addition, the mixture approach demonstrated that in order to minimize sample conditional risk in general, the joint a posteriori probability density of all parameters characterizing the mixture must be computed. Fralick,[2,14] for example, had to make the assumption that $F(B_i|\{X_s\}_1^{n-1}, B_j) = F(B_i|\{X_s\}_1^{n-1})$ because he did not compute joint densities.

Another contribution of the mixture approach is that classical results on Bayes estimates and maximum-likelihood estimates can be applied. It is shown in Chapter V that a uniqueness requirement, imposed by classical methods when obtaining the asymptotic variance of a maximum likelihood estimate, is replaced by the identifiability requirement when the c.d.f. is a mixture. Also in Chapter V, the asymptotic variances of parameters characterizing a binary, one dimensional, gaussian, non-supervisory problem are obtained. Previous investigators[4] had assumed all parameters known except the one being estimated-an assumptionwe do not make.

In addition to the development of the mixture approach, a class of non-supervisory problems is defined (Chapter 2). This class of nonsupervisory problems includes such problems as (a) any number of M pattern classes are possibly active causing each sample X_s, (b) the samples $\{X_s\}_{n-v+1}^{n}$ are not parameter conditionally independent, and (c) sets of samples are from the same pattern class with the pattern class unknown.

A second contribution is an application of histogram and empirical c.d.f. concepts to the nonsupervisory problem. When there is supervision, a histogram can always be obtained to approximate a class-conditional c.d.f. $F(X_s|\omega_i)$. In Chapter III it is shown that the use of a histogram to approximate a class-conditional c.d.f. $F(X_s|\omega_i)$, when the samples are not classified, results in a mixture of multinomial distributions. Whereas in the supervisory case parameters characterizing a multinomial distribution (histogram concept) can always be uniquely found, this is not true in the nonsupervisory case. It is shown in Chapter III that such parameters can be uniquely found, for the binary nonsupervisory case for example, if at least three samples from the same pattern class are taken at once (the class, of course, being unknown).

Whether the family has members with continuous functional forms or is multinomial, the mixture approach applies. The problem reduces to a classical problem of computing the a posteriori probability distribution of B if the objective is to minimize sample-conditional risk, or to finding a consistent estimator for B if the objective is suboptimum. In Chapter V, a consistent minimum distance estimator of B is given for a class of nonsupervisory problems where the classes of mixtures are identifiable.

CHAPTER II

NONSUPERVISION AND PARAMETER-CONDITIONAL MIXTURES

2.1 Mixtures and Parameter Conditional Mixtures

In this chapter a parameter-conditional mixture is defined. The type of the mixture depends upon the a priori knowledge used in its construction. By approaching nonsupervisory problems through first defining mixtures, we are able to define precisely different nonsupervisory problems and the a priori knowledge they utilize. For example, the mixture defined in Section 2.4 is used in Chapter III to apply histogram concepts to the nonsupervisory problem. The mixture defined in Section 2.2 corresponds to the nonsupervisory problem discussed in the Literature Survey. The mixture defined in Section 2.5 arises when more than one class can be active on the same sample. The minimum conditional risk solution given in Chapter IV applies to all the nonsupervisory problems discussed in this chapter.

A mixture results when a vector X can be partitioned W ways, $\pi_1, \pi_2, \ldots, \pi_W$. If, for example, $X = \{X_s\}_{n-v+1}^{n}$ with a single pattern class active causing each X_s, there are $W = M^v$ ways the pattern classes could be active to cause X. If, as another example, $X = X_s$ with a single pattern class active causing X_s, there are $W = M$ ways the pattern classes could be active to cause X. Since the partitions are mutually exclusive and exhaustive,

$$F(X) = \sum_{r=1}^{W} F(X|\pi_r)\, P(\pi_r) \qquad (2.1)$$

where F(X) is called the mixture c.d.f., $F(X|\pi_r)$ the rth partition-conditional c.d.f., and $P(\pi_r)$ the rth mixing parameter.

When we speak of a family of gaussian c.d.f.'s or a family of multinomial c.d.f.'s, we have in mind the nature of the parameters which characterize the family. It is therefore appropriate to define a parameter-conditional mixture c.d.f. $F(X|B)$ constructed using the family $\{F(X|\pi_r, B_r)\}$ of rth partition, parameter-conditional c.d.f.'s. To do this, define

$$B = B_1 \cup B_2 \cup \ldots \cup B_W \cup B_{W+1} \tag{2.2}$$

where

B_r, $r=1,2,\ldots,W$.: vector characterizing rth partition-conditional c.d.f.

$$B_{W+1} = \{P(\pi_r)\}_1^W \tag{2.3}$$

Thus B is simply the collection of the mixing parameters and all entries in $B_1,\ldots,B_W$. In other words, B contains all the fixed but unknown parameters characterizing the problem. Since (X, π_1), $(X,\pi_2),\ldots,(X,\pi_W)$ are mutually exclusive and exhaustive events,

$$\begin{aligned} F(X|B) &= \sum_{r=1}^{W} F(X,\pi_r|B) \\ &= \sum_{r=1}^{W} F(X|\pi_r,B)\, P(\pi_r|B) \end{aligned} \tag{2.4}$$

Now, the rth partition-conditional c.d.f. is characterized by B_r,

$$F(X|\pi_r,B) = F(X|\pi_r,B_r) \tag{2.5}$$

and since B contains $P(\pi_r)$,

$$P(\pi_r|B) = P(\pi_r) \tag{2.6}$$

Thus, (2.4) becomes

$$F(X|B) = \sum_{r=1}^{W} F(X|\pi_r,B_r)\, P(\pi_r) \tag{2.7}$$

If we are given $F(X)$, W, and the family $\{F(X|\pi_r, B_r\}$, then when can B be uniquely found? Or, put another way, given $F(X)$, when does $F(X) = F(X|B)$ have a unique solution B_o, which is the true value of B. The answer is that B_o can be uniquely found when the class of parameter-conditional mixtures is identifiable, sufficient conditions for which are given in Appendix A.

We now proceed to relate (2.7) to nonsupervisory problems arising in practice.

2.2 $X = X_s$ with Single Class Active

Let $X = X_s$ with one of M pattern classes possibly active. Then $W = M$ and (2.7) becomes

$$F(X_s|B) = \sum_{i=1}^{M} F(X_s|\omega_i,B_i)\, P_i \tag{2.8}$$

This parameter-conditional mixture (2.8) arises when samples $X_1,X_2,\ldots,X_n$ are parameter-conditionally independent.

2.3 v Samples Parameter-Conditionally Dependent

Let $X=\{X_s\}_{n-v+1}^{n}$ with a single pattern class active causing each sample X_s. Then $W = M^v$. Equation (2.7) becomes

$$F(\{X_s\}_{n-v+1}^{n}|B) = \sum_{r=1}^{M^v} F(\{X_s\}_{n-v+1}^{n}|\pi_r,B_r)\, P(\pi_r) \tag{2.9}$$

A mixture of this form arises when making a decision on sample X_n if X_n, given π_r and B_r, is statistically dependent on the previous (v-1) samples. The distribution function of X_n, conditioned on $\{X_s\}_{n-v+1}^{n-1}$ and B, can be expressed as

$$F(X_n|B,\{X_s\}_{n-v+1}^{n-1}) = \frac{\displaystyle\sum_{r=1}^{M^v} F(\{X_s\}_{n-v+1}^{n}|B_r^v,\, \pi_r^v)\, P(\pi_r^v)}{\displaystyle\sum_{r=1}^{M^{(v-1)}} F(\{X_s\}_{n-v+1}^{n-1}|B_r^{v-1},\pi_r^{v-1})\, P(\pi_r^{v-1})} \tag{2.10}$$

where π_r^v denotes the rth partition for samples $X_{n-v+1},\ldots,X_n$, and π_r^{v-1}

denotes the rth partition for samples $X_{n-v+1}, \ldots, X_{n-1}$.

Thus, when the v samples are statistically dependent, a priori knowledge must include the family $\{F(\{X_s\}_{n-v+1}^{n} | B_r, \pi_r)\}$ of multidimensional rth partition, parameter-conditional c.d.f.'s, the dimension of each member increasing as v increases. Furthermore, the number of terms in this mixture grows as v increases.

2.4 $X = X_s = \{X_{s_k}\}_1^v$ with Single Pattern Class Active

Let $X = X_s = X_{s_1}, X_{s_2}, \ldots, X_{s_v}$ with class ω_i active for all v samples. The parameter-conditional mixture c.d.f. $F(X_s|B)$ is

$$F(X_s|B) = \sum_{i=1}^{M} F(\{X_{s_k}\}_1^v | \omega_i, B_i)\, P_i \tag{2.11}$$

This mixture does not grow with increasing v as did the previous mixture because the statistically dependent samples are supervised. The a priori knowledge used to construct this mixture is knowledge of M, the family, and the fact that $X_s = \{X_{s_k}\}_1^v$ with one pattern class active for all samples.

We will find in Chapter III that this type of mixture arises when applying the histogram concept to nonsupervisory problems. By taking v samples at the sth observation with pattern class ω_i active, the class of mixtures may be identifiable whereas it would not be with only one sample taken.

2.5 $X = X_s$ with Interclass Interference

Let $X = X_s$ with any number of M classes possibly active causing X_s, a situation we will call interclass interference. The a priori knowledge also includes knowledge of M, the family, and that class ω_i is active on the sth sample with probabilitiy P_i. Since a class ω_i is either active or not for each sample X_s, there are 2^M mutually exclusive and exhaustive ways that the sth sample can occur. Thus the parameter-conditional mixture c.d.f. $F(X_s|B)$ is

$$F(X_s|B) = \sum_{r=1}^{2^M} F(X_s|\pi_r,B_r)\, P(\pi_r) \tag{2.12}$$

2.6 Two Possible Sets of Mixing Parameter

Let $X = X_s$ and a single class ω_i active for X_s. The a priori knowledge includes knowledge that $M = 2$, the family is known, and that there are two possible sets of mixing parameters defined as follows:

It is known that either P_1 or $(1-P_1)$ is equal to P; $P_1 = P$ with probability Q, and $(1-P_1) = P$ with probability $(1-Q)$, $Q = 0$ or 1. Since the events $P_1 = P$ and $(1-P_1) = P$ are mutually exclusive (assume $P \neq \frac{1}{2}$), the parameter conditional mixture c.d.f. is

$$\begin{aligned} F(X_s|B) = {} & Q\left[P\, F(X_s|\omega_1,B_1) + (1 - P)\, F(X_s|\omega_2,B_2)\right] + \\ & (1-Q)\left[(1-P)\, F(X_s|\omega_1,B_1) + P\, F(X_s|\omega_2,B_2)\right] \end{aligned} \tag{2.13}$$

where

$$B = (Q,\ P,\ B_1,\ B_2) \tag{2.14}$$

Define

$$F_1(X_s|B) = P\, F(X_s|\omega_1,B_1) + (1-P)\, F(X_s|\omega_2,B_2)$$

$$F_2(X_s|B) = (1-P)\, F(X_s|\omega_1,B_1) + P\, F(X_s|\omega_2,B_2)$$

Equation (2.14) then simplifies to

$$F(X_s|B) = Q\, F_1(X_s|B) + (1-Q)\, F_2(X_s|B) \tag{2.15}$$

As the problem is formulated, Q is either 1 or 0 since only one of the two sets of mixing parameters is active at a given time. Thus, (2.15) is a parameter-conditional mixture with one mixing parameter of value zero. The sufficient conditions given in Appendix A require all mixing parameters to be greater than zero but less than one. We therefore cannot conclude sufficient conditions for identifiability in this present problem. On the other hand, the fact that one of the mixing parameters has value P is a priori knowledge and should not impose greater constraints on the class of resulting parameter-conditional mixtures for identifiability. This shows the need for a study of identifiability when a mixture has one or more mixing parameters of value zero, and corresponds to the nonsupervisory problem with an unknown number of pattern classes M.

2.7 Given a Set of Families

Consider now a situation where there are R possible families, $\mathcal{F}_j = \{F^j(X_s|\omega_i,B_i)\}$, $j = 1,2,\ldots,R$. This might correspond to a problem where the class-conditional c.d.f. depends upon some parameter, for example phase, which changes from sample to sample, and takes on R possible values. Or, it might correspond to a problem where the noise statistics change from sample to sample, being represented by one of R possible c.d.f.'s. We will now assume that the samples are classified but that the families are not. That is, let $X = X_s$ with ω_i known active causing X_s and the jth family active with probability Q_j, $0 < Q_j < 1$, $j = 1,2,\ldots,R$. Then

$$F(X_s) = \sum_{j=1}^{R} Q_j \, F^j(X_s|\omega_i), \quad \omega_i \text{ known} \tag{2.16}$$

Thus the probability distribution of X_s is given by a mixture c.d.f. even though the samples are classified. In this case, the families active causing the samples are unclassified.

It is possible to give other examples where mixtures arise, by carefully defining the a priori knowledge available and using it to construct the mixture. We now proceed however, to Chapter III, where a construction technique is developed for approximating ith-class, parameter-conditional c.d.f.'s with multinomial distributions, utilizing available a priori knowledge about the c.d.f.'s. This, put another way, is the application of the histogram concept to nonsupervisory problems.

CHAPTER III

CONSTRUCTING ith CLASS, PARAMETER-CONDITIONAL C.D.F.'S

3.1 Introduction

Knowledge of the family or possible families of rth partition, parameter-conditional c.d.f.'s is required a priori knowledge in constructing the mixtures in Chapter II. The purpose of this chapter is to apply the histogram concept to nonsupervisory problems. To do this, we develope a construction method where multinomial c.d.f.'s are used to approximate ith class, parameter-conditional c.d.f.'s, utilizing available a priori knowledge about the c.d.f.'s.

3.2 The Fixed Bin Model

In general let $X_1 = \{X_{1_k}\}_1^{n_1}$, $X_2 = \{X_{2_k}\}_1^{n_2}, \ldots, X_n = \{X_{n_k}\}_1^{n_n}$ be n sequences of samples, the samples in sequence X_s coming from class ω_i. Although it is known that samples in a given sequence are from the same class, this class is unknown. In terms of the notation in the previous chapter, $X = X_s$ and $W = \mathbf{M}$.

Consider now the nonsupervisory problem where $n_1 = n_2 = \ldots = n_n = v$, the samples X_{s_k}, $k = 1,2,\ldots,v$, are parameter-conditionally independent, but for a given vector X_{s_k}, the different components are in general parameter-conditionally dependent.

X_{s_k} is an ℓ dimensional vector. We quantize each of these dimensions into R levels, obtaining R^ℓ, ℓ-dimensional "cubes" on "bins". Each ℓ dimensional bin has the same volume. X_{s_k} can lie in any of these R^ℓ bins, or in the $(R^\ell + 1)$st bin representing the remaining part of the ℓ-dimensional space. The bins are indexed and indicated by B_ξ, $\xi = 1,2,\ldots,(R^\ell + 1)$.

$F(X_s)$ is now approximated using the vector set $\underline{P}^o$ of fixed but unknown probabilities, $p_1^o, p_2^o, \ldots, p_{R^\ell}^o$, where p_ξ^o is the amount of probability in bin B_ξ of the sample space. Any probability in bin $B_{R^\ell+1}$, is given by

$$p_{R^\ell+1}^o = 1 - \sum_{\xi=1}^{R^\ell} p_\xi^o \tag{3.1}$$

In like manner, the ith class-conditional c.d.f., $F(X_s|\omega_i)$, is approximated by the vector set $\underline{P}^i$ of fixed but unknown probabilities, $p_1^i, p_2^i, \ldots, p_{R^\ell}^i$, where p_ξ^i is the amount of probability from $F(X_s|\omega_i)$, in bin B_ξ of the sample space.

Analogous to (3.1),

$$p_{R^\ell+1}^i = 1 - \sum_{\xi=1}^{R^\ell} p_\xi^i \tag{3.2}$$

The mixture corresponding to the nonsupervisory problem under consideration is of the same form as the mixture described in Section 2.4, since W = M and a single class is active for all v samples in a sequence. Therefore,

$$F(X_s|B_i) = \sum_{i=1}^{M} F(X_s|\omega_i, B_i)\, P_i \tag{3.3}$$

Under the framework of the approximations described above, (3.3) implies the following:

$$p_\xi^o = \sum_{i=1}^{M} p_\xi^i\, P_i, \qquad \xi = 1,2,\ldots,R^\ell + 1 \tag{3.4}$$

A binary (M = 2) one dimensional (ℓ = 1) example of this fixed bin model is shown in Fig. 1.

Since X_s is a sequence of v vector samples, samples fall in v of the $(R^\ell+1)$ bins of the sample space, not all bins being necessarily different. Let this relative frequency in the bins during the sth sequence be denoted by

$$V_s = (v_{s_1}, v_{s_2}, \ldots, v_{s_{R^\ell+1}}) \tag{3.5}$$

The distribution of V_s, given the class ω_i and $\underline{P}^i$, is

$$P(V_s|\omega_i,\underline{P}^i) = \frac{v!}{v_{s_1}!\ldots v_{s_{R^\ell+1}}!} \prod_{\xi=1}^{R^\ell+1} \left[p_\xi^i\right]^{v_{s_\xi}} \tag{3.6}$$

Approximating $F(X_s|\omega_i)$ by (3.6), we obtain the following parameter-conditional mixture of multinomial distributions:

$$P(V_s|B) = \sum_{i=1}^{M} P(V_s|\omega_i,B_i)\ P_i \tag{3.7}$$

where

$$B_i = \underline{P}^i$$

$$B = (\underline{P}^1,\ldots,\underline{P}^M,\ P_1,\ldots,P_M)$$

We now turn attention to the problem of estimating the bin probabilities corresponding to a single bin, say B_1. The M probabilities associated with bin B_1 are $p_1^1,\ p_1^2,\ldots,p_1^M$. For convenience, drop the subscript 1 in what follows. Then $(1-p^1)$, $(1-p^2),\ldots,(1-p^M)$ are the respective probabilities corresponding to all bins except B_1. What we have done here is the same as if we had originally chosen only two bins for the fixed bin model. The probability generating function of the binomial distribution $P(V_s|p^i,v,\omega_i)$ is $(p^iZ + 1 - p^i)^v = (1 + p^is)^v$ where $s = Z - 1$. Taking the probability generating function of both sides of (3.7), which is a mixture of binomial c.d.f.'s for this discussion, gives

$$\sum_{i=1}^{M} P_i(1 + s\ p^i)^v = (1 + s\ p^o)^v,\quad \text{all } s. \tag{3.8}$$

$$\sum_{i=1}^{M} P_i = 1$$

with $s = 1$, this is equivalent to

$$\sum_{j=0}^{v} \binom{v}{j}\left[\left\{\sum_{i=1}^{M} P_i\ (p^i)^j\ (1-p^i)^{v-j}\right\}-(p^o)^j\ (1-p^o)^{v-j}\right] = 0$$

$$\sum_{i=1}^{M} P_i = 1$$

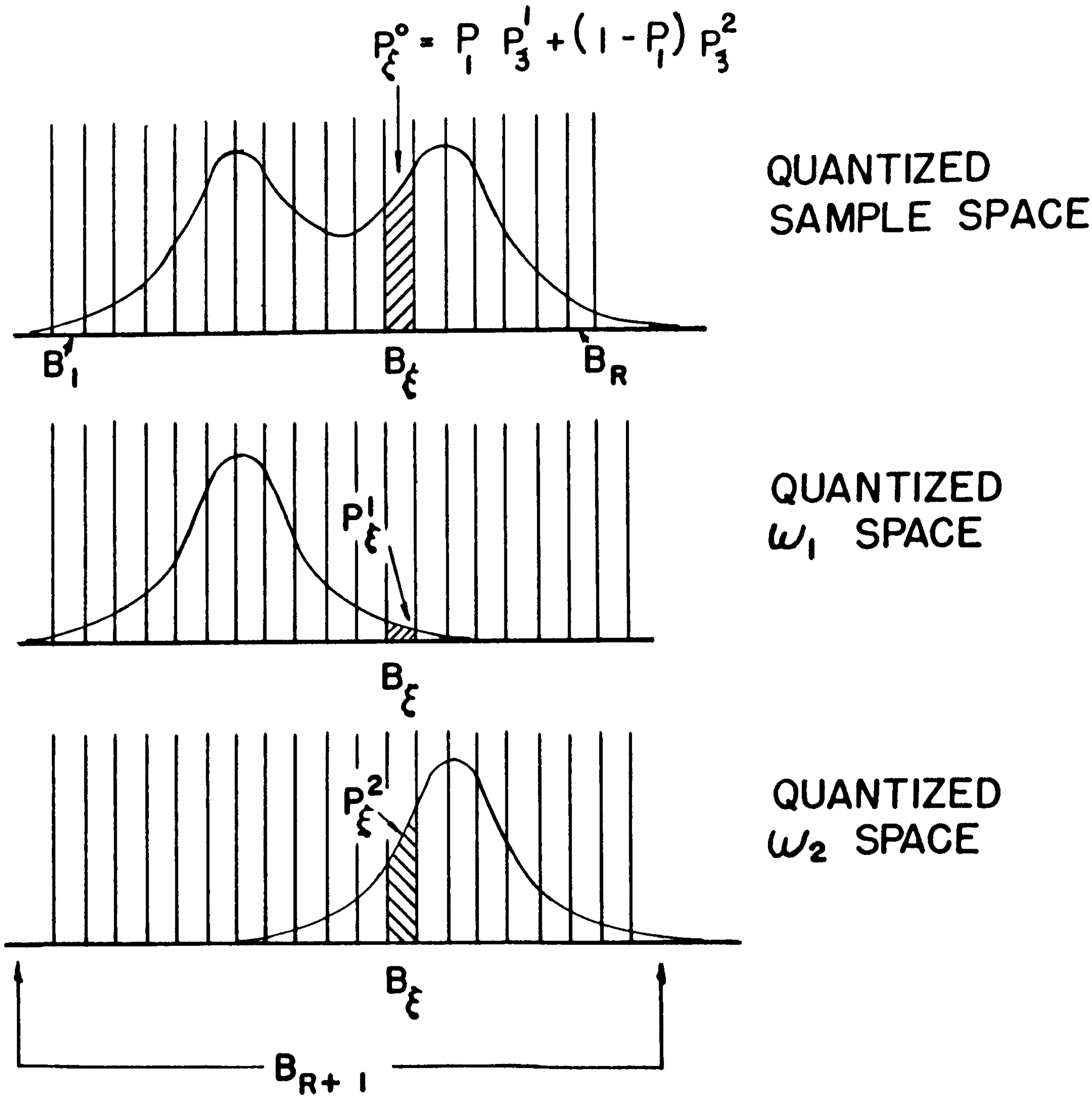

FIG. 1 - QUANTIZED SPACES

or

$$\sum_{i=1}^{M} P_i(p^i)^j\ (1-p^i)^{v-j} = (p^o)^j\ (1-p^o)^{v-j},\ j = 0,1,\ldots,v$$

$$\sum_{i=1}^{M} P_i = 1 \tag{3.9}$$

Proposition A.5 in Appendix A guarantees a unique solution of (3.9) for $p^1,p^2,\ldots,p^M$, P_1, $P_2,\ldots,P_M$, given the right side of (3.9), if $v \geq 2M-1$.

The significance of the above result for engineering purposes is that a priori knowledge sufficient to solve such a nonsupervisory problem, where the form of the statistics is unknown, is provided by the existence of sequences of v samples from the same pattern class. A binary (M = 2) one dimensional ($\ell = 1$) example will help to illustrate (3.9). For this binary case it is sufficient that v = 3; i.e., three samples be taken at a time with the same class active. Then V_s can occur four ways: three occurrences in bin B_1, two occurrences in bin B_1, one occurence in bin B_1, or 0 occurrences in bin B_1. These relative frequencies are consistent estimators of $(p^o)^j(1-p^o)^{v-j}$, j = 0,1,2,3, respectively. Then (3.9) can be solved for estimators of p^1, p^2, and P_1 in terms of these consistent estimators.

Another way to obtain estimators of p^1, p^2, and P_1 is given by Blischkee,[11] who derived moment estimators. Such moment estimators can be substituted into the decision equation developed in the next chapter, thereby obtaining a suboptimum solution of this nonsupervisory problem.

3.3 Utilizing Additional A Priori Knowledge about the C.D.F.'s

If it is known, for example, that a c.d.f. is symmetrical, then approximating this c.d.f. by a multinomial c.d.f. does not utilize the symmetrical knowledge. For this case we would use an appropriately defined symmetrical multinomial distribution to approximate the c.d.f. If, as another example, it is known that the ith class, parameter-conditional c.d.f.'s differ only by translational parameters, we would approximate each c.d.f. by an appropriately defined translated multinomial c.d.f.

We have not yet said how we propose to count the bin probabilities in ℓ dimensional space, although writing $p_1^i, p_2^i, \ldots, p_{R^\ell}^i$ indicates we must have had some counting procedure in mind. One method of counting is to redefine the bin probabilities as

$$p^i_{j_1, j_2, j_3, \ldots, j_\ell} \qquad 1 \leq j_a \leq R \quad \text{for all } a. \tag{3.10}$$

Then let

$$\begin{aligned} p_1^i &= p^i_{1,1,\ldots,1} \\ &\vdots \\ p_R^i &= p^i_{R,1,\ldots,1} \\ p_{R+1}^i &= p^i_{R,2,1,\ldots,1} \end{aligned} \tag{3.11}$$

etc.

It is convenient to define a vector ψ_ξ whose entries are the subscripts corresponding to quantum level ξ. Then, (3.11) becomes

$$p_\xi^i = p_{\psi_\xi}^i, \quad \xi = 1, 2, \ldots, R^\ell \tag{3.12}$$

Define the family of ith class, parameter-conditional multinomial c.d.f.'s where $B_i = \underline{P}^i$ by $\mathcal{F}_p = \{F(X|\omega_i, \underline{P}^i)\}$. This is the family used in the construction of the parameter-conditional mixture (3.7). Next define the family $\mathcal{F}_{TP}$ of multinomial c.d.f.'s differing only by translational vectors, $\{\theta_i\}$. To accomplish this, define $\underline{P}_{\theta_o}$ where θ_o is a vector of ℓ indices

$$\theta_o = \left(\frac{R+1}{2}, \frac{R+1}{2}, \ldots, \frac{R+1}{2}\right), \quad R \text{ odd}$$

The vector θ_o locates the center bin in the ℓ dimensional space with R^ℓ quantum levels used for representing $\underline{P}_{\theta_o}$. In terms of $\underline{P}_{\theta_o}$, the vector $\underline{P}^i$ characterizing the ith class-conditional c.d.f. is expressed as

$$\underline{P}^i = \underline{P}_{\theta_o - \theta_i}, \quad p^i_{R^\ell+1} = 0, \quad i = 1,2,\ldots,M \tag{3.13}$$

Also define a family $\mathcal{F}_{STP}$ of symmetrical multinomial c.d.f.'s differing only by translational vectors, $\{\theta_i\}$, by letting $\underline{P}_{S\theta_o}$ be a vector whose entries are symmetrical in each of the ℓ dimensions.

Returning to the nonsupervisory problem under consideration, assume it is known that the ith class-conditional c.d.f.'s are all identical except for different translational parameters. We approximate these c.d.f.'s by members of the family $\mathcal{F}_{TP}$. The distribtuion of X_s is then approximated by the parameter-conditional mixture c.d.f.,

$$P(V_s|B) = \sum_{i-1}^{M} P(V_s|\omega_i,\underline{P}_{\theta_o},\theta_i,\theta_o)\, P_i, \quad v \geq 1 \tag{3.14}$$

where

$$B_i = (\underline{P}_{\theta_o},\ \theta_i,\ \theta_o)$$

$$B = (\underline{P}_{\theta_o},\ \theta_1,\ldots,\theta_M,P_1,\ldots,P_M) \tag{3.15}$$

The a priori knowledge that the ith class-conditional c.d.f.'s differ only by translational vectors reduced the number of entries in the vector B, characterizing the mixture, by $(M-1)\,R^\ell - M$ which may be a considerable reduction. If the family $\in \mathcal{F}_{STP}$, instead of $\mathcal{F}_{TP}$, the number of parameters characterizing the mixture is further reduced by $\left[\left(\frac{R+1}{2}\right) - 1\right]^\ell$,

Another way to consider a symmetrical multinomial c.d.f. is as follows: Let

$$F(-x-\theta_i|\omega_i,\theta_i) = F(x-\theta_i|\omega_i,\theta_i), \quad \ell = 1, \quad v \geq 1 \tag{3.16}$$

be approximated by a symmetrical multinomial c.d.f., where θ_i is a translational parameter. If $x_1, x_2, \ldots, x_n$ are samples from $F(x|\omega_i, \theta_i)$, then samples $\theta_i - (x_1-\theta_i)$, $\theta_i - (x_2-\theta_i), \ldots, \theta_i-(x_n-\theta_i)$ are just as likely to have occurred. That is, given symmetry, θ_i, and n samples, we really have 2n samples as far as constructing the c.d.f. is concerned. We might write

$$\{x_s\}_1^n \cup \{\begin{matrix}\text{knowledge of symmetry}\\ \text{and } \theta_i\end{matrix}\} = \{x_s\}_1^n \cup \{2\theta_i - x_s\}_1^n$$

Or defining a "symmetry operator" S_i,

$$S_i\ (\{x_s\}_1^n,\ \theta_i) = \{x_s\}_1^n \cup \{2\theta_i - x_s\}_1^n \tag{3.17}$$

So, **knowledge** of symmetry and θ_i maps the n received samples to 2n samples. If it is also known that the ith class, parameter-conditional c.d.f.'s differ only by translational parameters, and say n samples are received from each of M classes, then there are 2n M samples available for the construction of each ith class, parameter-conditional c.d.f. It is obvious that such a priori knowledge increases convergence rate if the system converges.

3.4 Family of Multinomial C.D.F.'s with Spacial Constraints and v = 1

Let $x = x_s$, a single sample, $W = M = 2$, $\ell = 1$, and assume it known a priori that $F(x|\omega_2) = 0$ for $x \leq \theta_1$ and $F(x|\omega_1) = 1$ for $x \geq \theta_2$ where θ_1 and θ_2 are translational parameters. This latter constraint corresponds to an approximation that can be made when the "signal-to-noise" ratio is "sufficiently large," and each class-conditional c.d.f. is symmetrical about its translation parameter.

Samples x_s, $s = 1,2,\ldots,n$, which fall $\leq \theta_1$, given θ_1, are thus known to have been caused by class ω_1; and samples which fall $\geq \theta_2$, are known to have been caused by class ω_2. Only samples greater than θ_1 and less than θ_2, given θ_1 and θ_2, are not classified.

Define $\{x_{s_n}\}_1^{n_1}$ as those samples $\leq \theta_1$, and $\{x_{s_n}\}_{n_1+1}^{n_2}$ as those samples $\geq \theta_2$, and $\{x_{s_n}\}_{n_2+1}^{n_3}$ as those samples greater than θ_1 and less than θ_2. Let D_1 be an operator mapping $(\{x_s\}_1^n, \theta_1)$ to $\{x_{s_n}\}_1^{n_1}$ and S_1 an operator mapping $\{x_{s_n}\}_1^{n_1}$ to $\{x_{s_n}\}_1^{2n_1}$ as described by (3.17). Similarly, let S_2 D_2 and S_3 D_3 be the respective operators for the samples $n_1+1,\ldots,n_2$, and the samples $n_2+1,\ldots,n_3$.

Since a single class is active causing each sample, the parameter-conditional mixture is of the form (2.8) for any of the samples $n_2+1,\ldots,n_3$, but is a degnerate mixture for any of the samples $1,\ldots,n_1$, or $n_1+1,\ldots,n_2$:

$$F(V_s|B) = \begin{cases} F(V_s|\omega_1,B_1), & D_1x_s \leq \theta_1 \\ F(V_s|\omega_2,B_2), & D_2x_s \geq \theta_2 \\ \sum_{i=1}^{2} F(V_s|\omega_i,B_i)\, P_i, & \theta_1 < D_3x_s < \theta_2 \end{cases} \tag{3.18}$$

where

$$\begin{aligned} B_1 &= (\underline{P}^1,\ \theta_1) \\ B_2 &= (\underline{P}^2,\ \theta_2) \\ B &= (B_1, B_2, P_1), \quad v = 1 \text{ and} \end{aligned} \tag{3.19}$$

V_s is the relative frequency in the bins resulting from the application of the symmetry operator to x_s.

An example where the samples $n_2+1,\ldots,n_3$ were not used is given in reference 7; the system objective considered there was to minimize sample-conditional probability or error. Setting $P_{1_o} = \frac{1}{2}$ and known, θ_{1_o} and θ_{2_o} unknown, it is shown that the system converges when the ith-class, parameter-conditional c.d.f.'s were guassian, but not known a priori. Equation (3.18), however, shows how to use the samples $n_2+1,\ldots,n_3$ in constructing the mixture. The minimum sample-conditional probability of error solution given in Chapter IV, utilizing the mixture (3.18), gives an optimum solution to this problem reported in reference 7.

Rather than using the symmetry operator we can define the symmetrical vector $\underline{P}_s^i$. Then, the form of (3.18) becomes

$$F(V_s|B) = \begin{cases} F(V_s|\omega_1,B_1) & D_1x_s \leq \theta_1 \\ F(V_s|\omega_2,B_2) & D_2x_s \geq \theta_2 \\ \sum_{i=1}^{2} F(V_s|\omega_i,B_i)\, P_i & \theta_1 < D_3x_s < \theta_2 \end{cases}$$

where

$$B_1 = (\underline{P}_s^1, \theta_1)$$
$$B_2 = (\underline{P}_s^2, \theta_2)$$
$$B = (B_1, B_2, P_1)$$

with $\underline{P}_s^i$ the vector $\underline{P}^i$ with symmetry about its middle entry. V_s is the actual relative frequency in the bins resulting from sample x_s.

If it is also known that the ith class, parameter-conditional c.d.f.'s differ only by translational parameters, then

$$B_1 = (\underline{P}_{s\theta_o}, \theta_1)$$
$$B_2 = (\underline{P}_{s\theta_o}, \theta_2)$$
$$B = (\underline{P}_{s\theta_o}, \theta_1, \theta_2, P_1)$$

with $\underline{P}_{s\theta_o}$ the vector common to both classes and symmetrical about θ_o.

3.5 <u>Families of Empirical C.D.F.'s</u>

In the rest of this chapter we consider the problem of applying empirical c.d.f. concepts to a nonsupervisory problem with $\ell = 1$. A single, one dimensional sample x_s is taken at the sth observation. For the remainder of this chapter, let $(x_{n_1},\ldots,x_{n_n})$ be the ordered samples of n one dimensional samples from the c.d.f. F(x). Denote the corresponding empirical c.d.f. by $F_n(x) = F(x|\{x_{n_s}\}_1^n)$. $F_n(x)$ is constructed from the ordered samples as follows:

$$F_n(x) = \begin{cases} 0 & x < x_{n_1} \\ \frac{\gamma-1}{n} & x_{n_{\gamma-1}} \leq x < x_{n_\gamma}, \quad \gamma = 1,2,\ldots,n \\ 1 & x \geq x_{n_n} \end{cases} \tag{3.20}$$

The parameters characterizing $F_n(x)$ are the n ordered samples $\{x_{n_s}\}_1^n$, the number growing as n increases.

Let a single class be active causing each sample; then there are $W = M^n$ ways the samples could have been caused. The distribution of x given $\{x_{n_s}\}_1^n$ can be expressed as

$$F(x|\{x_{n_s}\}_1^n,\{P(\pi_r)\}) = \sum_{r=1}^{M^n} F(x|\pi_r,\{x_{n_s}\}_1^n)\ P(\pi_r) \tag{3.21}$$

Although (3.21) may appear to be a parameter conditional mixture, this is not the case since the parameters characterizing this c.d.f. are random variables, which grow in number with increasing n. The Adaptive Bin Model is now introduced to provide an engineering solution to this difficulty.

3.6 Adaptive Bin Model

A model is next obtained where the R bins are R coverages[13] formed from the ordered samples. There are numerous ways that coverages[13,17,18] can be formed given a sequence of ordered samples $\{x_{n_s}\}_1^n$ from a c.d.f. F(x). We will consider one such way to be used in an adaptive bin model. This model involves an approximation which improves as n increases.

Let the number of samples be $n = Rv - 1$. If $x_{(Rv-1)_v}$ is the vth smallest sample, then it is well known[13,17] that

$$F(x_{(Rv-1)_v}) \xrightarrow[v \to \infty]{P} \frac{1}{R} \tag{3.22}$$

$x_{(Rv-1)_v}$ is called the 1st sample quantile ($\frac{1}{R}$th quantile) and $F(x_{(Rv-1)v})$ the corresponding 1st population quantile. The adaptive bin model is established by defining the following:

$n = Rv - 1$ = number of one dimensional samples

R = number of coverages (or adaptive bins)

$x_{n_0} = -\infty$

$x_{n_{n+1}} = +\infty$

v = number of samples in a bin

and the locations of the R adaptive bins are denoted by

$$\begin{aligned}&(x_{n_0}, x_{n_v}], (x_{n_v}, x_{n_{2v}}], \ldots, (x_{n_{(R-1)v}}, x_{n_{n+1}}) = \\ &\{(x_{n_{(\xi-1)v}}, x_{n_{\xi v}}]\}_{\xi=1}^{R-1}, (x_{n_{(R-1)v}}, x_{n_{n+1}})\end{aligned} \tag{3.24}$$

$F(x_{n_{\xi v}})$ is the amount of probability from the population c.d.f. in the interval $(x_{n_0}, x_{n_{\xi v}})$. It is well known[13,17] that

$$E[F(x_{n_{\xi v}})] = \frac{\xi v}{(Rv-1) + 1} = \frac{\xi}{R}, \quad \xi = 1,2,\ldots,R \tag{3.25}$$

The difference $U_\xi = [F(x_{n_{\xi v}})] - [F(x_{n_{(\xi-1)v}})]$ is called <u>converage ξ</u> corresponding to adaptive bin B_ξ. Using (3.25),

$$E[U_\xi] = \frac{(\xi v-(\xi-1)v)}{Rv} = \frac{1}{R}, \quad \xi = 1,2,\ldots,R \tag{3.26}$$

Thus the expected amount of probability is $1/R$ in all R adaptive bins. Furthermore, the adaptive bins (3.24) converge in probability to the intervals corresponding to the 1/Rth quantiles of the population c.d.f.[13,17] The adaptive bins thus become statistically stable in location as n becomes large.

Let $\underline{P}^o$ be the vector of bin probabilities characterizing the multinomial c.d.f. approximating F(x) as in the fixed bin model except that there is no (R+1)st. bin. Set $p_1^o = p_2^o = \ldots = p_R^o = \frac{1}{R}$, and let the adaptive bins (3.24) be approximations to the actual 1/Rth intervals of the c.d.f. F(x). Let $\underline{P}^i$ be the vector characterizing the c.d.f. $F(x|\omega_i)$, with p_ξ^i the amount of probability from the ith class in the actual ξth, 1/Rth interval of the c.d.f. F(x).

Another feature of the adaptive bin model is that coverages are used to approximate the R equal probability intervals of F(x), and the fixed bin model is then applied. The parameter-conditional mixture looks the same as for the fixed bin model:

$$F(x_s|B) = \sum_{r=1}^{M} F(x_s|B_i,\omega_i)\ P_i \tag{3.27}$$

except that

$$B_i = (\underline{P}^i,\ \{(x_{n_{(\xi-1)v}},\ x_{n_{\xi v}}]\}_{\xi=1}^{R} \tag{3.28}$$

That is, the parameters characterizing $F(x|\omega_i)$ include the locations of the adaptive bins.

The practical advantage of adaptive bins is that the bins are automatically placed where there are samples. That is, there are few bins where there are few samples and many bins where there are many samples.

CHAPTER IV

MINIMUM CONDITIONAL RISK SOLUTION FOR NONSUPERVISORY PROBLEMS

4.1 Optimum System Objective

We are interested in observing sample $X_n(v=1)$ and deciding which class ω_i is active or in observing the sequence $X_n = \{X_{n_k}\}_1^v$ of v samples and deciding which class is active.

For each $\omega_i \in \Omega$ it is possible to use any decision function $d \in D$. Let $L(d(X_n)|\omega_i)$, independent of B and $\{X_s\}_1^{n-1}$, be the class-conditional loss function defined at every point in the product space D X Ω. For any $d \in D$ and $\omega_i \in \Omega$, the class-conditional risk function $r(d|\omega_i)$ is defined as the average of the class-conditional loss function over the sample space:

$$r(d|\omega_i) = E[L(d(X_n)|\omega_i)] = \int L(d(X_n)|\omega_i)f(X_n|\omega_i)dX_n \tag{4.1}$$

For given decision function d, mixing parameters $\{P_i\}_1^M$, and vectors $\{B_i\}_1^M$, the parameter-conditional risk averaged over Ω is

$$r(d|B) = \sum_{i=1}^{M} r(d|\omega_i,B_i)P_i = \sum_{i=1}^{M} [\int L(d(X_n)|\omega_i)f(X_n|\omega_i,B_i)dX_n]P_i \tag{4.2}$$

Let $f(B|\{X_s\}_1^{n-1})$ be the sample conditional density of B, given the samples $\{X_s\}_1^{n-1}$, which will be computed shortly. Then the sample-conditional risk is

$$r(d|\{X_s\}_1^{n-1}) = \int r(d|B)\ f(B|\{X_s\}_1^{n-1})\ dB \tag{4.3}$$

since $r(d|B)$ is completely characterized by B.

Since $L(d(X_n|\omega_i)$ is independent of B and $\{X_s\}_1^{n-1}$, (4.3) can be written

$$r(d|\{X_s\}_1^{n-1}) =$$

$$\int d\ B\{\int d\ X_n[\sum_{i=1}^{M} L(d(X_n)|\omega_i)f(X_n|\omega_i,B_i)P_i]\}\ f(B|\{X_s\}_1^{n-1}) \tag{4.4}$$

Thus to minimize sample-conditional risk against a priori knowledge which includes a set of loss functions, the family of ith-class, parameter-conditional c.d.f.'s, M, and $f(B|\{X_s\}_1^{n-1})$, solve (4.4) for the decision function d.

If the loss function is a 0,1 loss function, X_n is a discrete random vector, and d is chosen to be the following:

$d(X_n)$: choose ω_j such that

$$P(X_n,\omega_j|\{X_s\}_1^{n-1}) = \sup_i \{P(X_n,\omega_i|\{X_s\}_1^{n-1})\}_{i=1}^M \tag{4.5}$$

then sample-conditional probability of error is minimized. When $f(X_n,\omega_i|\{X_s\}_1^{n-1})$ is continuous in X_n, the decision equation (4.4) with 0,1 loss function is equivalent to

$d(X_n)$: choose ω_j such that

$$f(X_n,\omega_j|\{X_s\}_1^{n-1}) = \sup_i \{f(X_n,\omega_i|\{X_s\}_1^{n-1})\}_{i=1}^M \tag{4.6}$$

4.2 Computation of $f(B|\{X_s\}_1^{n-1})$ for Mixtures

In order to minimize sample-conditional risk, $f(B|\{X_s\}_1^{n-1})$ must be computed where the following a priori knowledge is available:

(a) The family of ith class, parameter-conditional c.d.f.'s and M are known, and the parameter-conditional mixture c.d.f. $F(X_{n-1}|B,\{X_s\}_1^{n-2})$ thus constructed.

(b) Additional constraints on X or B to insure the class of mixtures is identifiable.

(c) F(B) - at least an appropriately defined uniform c.d.f., not ruling out the true value of B.

Working with density functions rather than c.d.f.'s, $f(B|\{X_s\}_1^{n-1})$ is given by Bayes Theorem as follows:

$$f(B|\{X_s\}_1^{n-1}) = \frac{f(X_{n-1}|B,\{X_s\}_1^{n-2})\ f(B|\{X_s\}_1^{n-2})}{f(X_{n-1}|\{X_s\}_1^{n-2})} \tag{4.7}$$

The demominator on the right side of (4.7) is a normalization constant which assures that $f(B|\{X_s\}_1^{n-1})$ integrates over the B space to unity. $f(B|\{X_s\}_1^{n-2})$ is the density in the B space at the (n-2) stage. $f(X_{n-1}|B,\{X_s\}_1^{n-2})$ is a function directly utilizing the a priori knowledge above. If, for example, the samples are parameter-conditionally independent, $f(X_{n-1}|B,\{X_s\}_1^{n-2})$ is given by (2.8). If the samples are parameter-conditonally dependent only on the last v samples, the form of $f(X_{n-1}|B,\{X_s\}_1^{n-2})$ is given by (2.10). If the samples are from multinomial distributions with v samples taken at the sth observation with a single class active as in the fixed bin model, X_{n-1} is a sequence of v samples and (2.11) is used.

If the identifiability requirement (b) assures the existence of an estimator for B_o (the true value of B) converging to B_o with probability one, then $f(B|\{X_s\}_1^{n-1})$ converges to a Dirac delta function[20] at B_o. In Chapter III we showed sufficient conditions for the existence of such an estimator for families of multinomial c.d.f.'s when using the fixed bin model. We will show in Chapter V (Theorem 5.1) that for any parameter-conditional mixture c.d.f. $F(X|B)$, the class of which is identifiable, continuous in X and B, such a consistent estimator for B_o exists.

For convenience we now limit considerations to the vector samples being parameter-conditionally independent. $F(X_{n-1}|B,\{X_s\}_1^{n-2})$ then has the form (2.8) such that (4.7) becomes

$$f(B|\{X_s\}_1^{n-1}) = \frac{\left[\sum_{i=1}^{M} f(X_{n-1}|\omega_i,B_i)P_i\right] f(B|\{X_s\}_1^{n-2})}{f(X_{n-1}|\{X_s\}_1^{n-2})} \tag{4.8}$$

Equation (4.8) is the fundamental result for the a posteriori probability density of the vector B characterizing a parameter-conditional mixture. It is used in the minimum sample-conditional risk equation (4.4). Sometimes it is desirable to obtain the a posteriori probability of just one parameter in B; for example, the Bayes estimate of such a parameter may be desired. Therefore, let γ_{K_j} be some parameter in B_K. The sample-conditional density of γ_{K_j} is obtained by integrating (4.8) with respect to all parameters in B not equal to γ_{K_j}. Integrating (4.8) in this fashion gives

$$f(\gamma_{K_j} | \{X_s\}_1^{n-1}) = \frac{\left[\sum\limits_{i \neq K} \int P_i \, f(X_{n-1} | B_i, \omega_i) \, f(B | \{X_s\}_1^{n-2}) \, d\bar{B} \right]}{f(X_{n-1} | \{X_s\}_1^{n-2})} \tag{4.9}$$

$$+ \frac{\int P_K \, f(X_{n-1} | B_K, \omega_K) \, f(B | \{X_s\}_1^{n-2}) \, d\bar{B}}{f(X_{n-1} | \{X_s\}_1^{n-2})}$$

where $\bar{B}$ is defined as the vector not containing parameter γ_{K_j} but containing all other parameters of B. Continuing with (4.9) we obtain

$$f(\gamma_{K_j} | \{X_s\}_1^{n-1}) =$$

$$\frac{\left[\sum\limits_{i \neq K} \int P_i \, f(X_{n-1} | B_i, \omega_i) \, f(\bar{B} | \{X_s\}_1^{n-2}, \gamma_{K_j}) \, d\bar{B} \right]}{f(X_{n-1} | \{X_s\}_1^{n-2})} f(\gamma_{K_j} | \{X_s\}_1^{n-2})$$

$$+ \frac{\left[\int P_K \, f(X_{n-1} | B_K, \omega_K) \, f(\bar{B} | \{X_s\}_1^{n-2}, \gamma_{K_j}) \, d\bar{B} \right]}{f(X_{n-1} | \{X_s\}_1^{n-2})} f(\gamma_{K_j} | \{X_s\}_1^{n-2}) \tag{4.10}$$

and since $f(X_{n-1}|B_i,\omega_i) = f(X_{n-1}|B_i,\omega_i,\gamma_{K_j},\{X_s\}_1^{n-2})$ and $P_K f(X_{n-1}|B_K,\omega_K) = f(X_{n-1},\omega_K|\overline{B},\gamma_{K_j},\{X_s\}_{n-2})$ where $\overline{B}$ is a vector containing all entries in B except γ_{K_j}, (4.10) becomes

$$f(\gamma_{K_j}|\{X_s\}_1^{n-1}) = \left[\sum \frac{f(X_{n-1},\omega_i|\{X_s\}_1^{n-2},\gamma_{K_j})}{f(X_{n-1}|\{X_s\}_1^{n-2})}\right] f(\gamma_{K_j}|\{X_s\}_1^{n-2}) \tag{4.11}$$

$$+ \frac{f(X_{n-1},\omega_K|\{X_s\}_1^{n-2},\gamma_{K_j})}{f(X_{n-1}|\{X_s\}_1^{n-2})} \frac{E\left[f(X_{n-1},\omega_K|\gamma_{K_j},\{X_s\}_1^{n-2})\right]}{f(X_{n-1},\omega_K|\{X_s\}_1^{n-2},\gamma_{K_j})} F(\gamma_{K_j}|\{X_s\}_1^{n-2})$$

where the expectation is a conditional expectation, conditioned on $\{X_s\}_1^{n-2}$ and γ_{K_j}, and taken with respect to $\overline{B}$. That is,

$$E\left[f(X_{n-1},\omega_K|\gamma_{K_j},\{X_s\}_1^{n-2}\right] =$$

$$\int f(X_{n-1},\omega_K|\overline{B}_K,\gamma_{K_j})\, f(\overline{B}|\gamma_{K_j},\{X_s\}_1^{n-2})\, d\overline{B} \tag{4.12}$$

Define the "weighting coefficients" within { } (4.11) by $C_i(\gamma_{K_j})$:

$$C_i(\gamma_{K_j}) = \frac{f(X_{n-1},\omega_i|\{X_s\}_1^{n-2},\gamma_{K_j})}{f(X_{n-1}|\{X_s\}_1^{n-2})} \tag{4.13}$$

Using these "weighting coefficients," (4.13) becomes

$$f(\gamma_{K_j}|\{X_s\}_1^{n-1}) = \Bigg[\sum_{i\neq K} C_i(\gamma_{K_j}) +$$

$$C_K(\gamma_{K_j}) \frac{E\left[f(X_{n-1},\omega_K|\gamma_{K_j},\{X_s\}_1^{n-2})\right]}{f(X_{n-1},\omega_K|\gamma_{K_j},\{X_s\}_1^{n-2})}\Bigg] f(\gamma_{K_j}|\{X_s\}_1^{n-2}) \tag{4.14}$$

The interpretation of (4.14) is as follows:

a) $\sum_{i\neq K} C_i(\gamma_{K_j})$ is the probability, conditioned on γ_{K_j} and $\{X_s\}_1^{n-2}$, that class ω_K was not active to produce the sample X_{n-1}. With probability $\sum_{i\neq K} C_i(\gamma_{K_j})$, the conditional density of γ_{K_j} at the (n-2) stage is thus retained.

b) $C_K(\gamma_{K_j})$ is the probability, conditioned on γ_{K_j} and $\{X_s\}_1^{n-2}$, that class ω_K was active to produce sample X_n. With probability C_K, the conditional density of γ_{K_j} at the (n-2) stage is updated in a supervisory manner. That is, if it is known X_{n-1} came from class ω_K, then (4.14) becomes

$$f(\gamma_{K_j}|\{X_s\}_1^{n-1}) = \frac{E\left[f(X_{n-1},\omega_K|\gamma_{K_j},\{X_s\}_1^{n-2})\right]}{f(X_{n-1},\omega_K|\gamma_{K_j},\{X_s\}_1^{n-2})} f(\gamma_{K_j}|\{X_s\}_1^{n-2}) \tag{4.15}$$

c) $E\left[f(X_{n-1},\omega_K|\gamma_{K_j},\{X_s\}_1^{n-2}\right]$ is involved in (4.14) because $f(X_{n-1},\omega_K|B_K)$ is in general a function of parameters other than γ_{K_j}.

4.3 Systems Minimizing Sample-Conditional Probability of Error

In this section we consider the design of systems minimizing sample-conditional probability of error. When $F(X_s|\omega_i)$ is approximated by a multinomial c.d.f. using the fixed bin model, X_s is a discrete random vector. We therefore use decision equation (4.5) with $P(X_n,\omega_i|\{X_s\}_1^{n-1})$ computed in terms of $f(B|\{X_s\}_1^{n-1})$ as follows:

$$P(X_n,\omega_i|\{X_s\}_1^{n-1}) = \int P(X_n,\omega_i|B,\{X_s\}_1^{n-1})\ f(B|\{X_s\}_1^{n-1})\ dB,\quad i = 1,2,\ldots,M$$

Since the samples $\{X_s\}_1^n$ are assumed parameter-conditionally independent,

$$P(X_n,\omega_i|\{X_s\}_1^{n-1}) = \int \left[P(X_n|\omega_i,B)P_i\right] f(B|\{X_s\}_1^{n-1})\ dB \quad i = 1,2,\ldots,M \tag{4.16}$$

with

$$B_i = \underline{P}^i,\ v \text{ known} \tag{4.17}$$

where in general X_s is a sequence of v samples as described in Chapter III.

Denote the bins that the v samples on the nth observation fall in by B_{η_K}, K = 1,2,...,v. Using this notation and (3.6) in (4.16),

$$P(X_n,\omega_i|\{X_s\}_1^{n-1}) = \frac{v!}{v_{\eta_1}!\dots v_{\eta_v}!}\int\left[\left\{\prod_{K=1}^{v} p_{\eta_K}^{i}\right\}P_i\right] f(B|\{X_s\}_1^{n-1})\, dB \qquad (4.18)$$

where v_{η_K} is one.

It is convenient to define the sample conditional expectation of $\left\{\frac{v}{v_{\eta_1}!\dots v_{\eta_v}!}\prod_{K=1}^{v} p_{\eta_K}^{i}\right\}P_i$ by $\left[p_\eta^i\right]_{n-1}$; that is,

$$\left[p_\eta^i\right]_{n-1} = P(X_n,\omega_i|\{X_s\}_1^{n-1}) = E\left[\left\{\frac{v!}{v_{\eta_1}!\dots v_{\eta_v}!}\prod_{K=1}^{v} p_{\eta_K}^{i}\right\}P_i|\{X_s\}_1^{n-1}\right] \qquad (4.19)$$

If v = 1, (4.19) reduces to

$$\left[p_\eta^i\right]_{n-1} = E\left[p_\eta^i\ P_i|\{X_s\}_1^{n-1}\right] \qquad (4.20)$$

when p_η^i is the probability from the ith class in the single bin in which sample X_n fell. Equation (4.20) used along with the decision equation (4.5) has an interesting interpretation: To minimize the sample-conditional probability of error when v = 1, while making a decision on the nth. sample, observe the bin into which the nth sample fell, say B_η. Then compute the expected amount of probability in bin B_η for all M classes and make decisions as follows: choose ω_j ∋

$$\left[p_\eta^j\right]_{n-1} = \sup_i \left\{\left[p_\eta^i\right]_{n-1}\right\} \qquad (4.21)$$

If v > 1, one observes the expected values (for each i) as in the right side of (4.19). These expected values do not have the simple interpretation as when v = 1.

Equation (4.7) requires $f(X_{n-1}|\omega_i,B_i)$ which is computed from $P(V_{n-1}|B_i,\omega_i)$ by

$$f(X_{n-1}|B_i,\omega_i) = K\ P(V_{n-1}|B_i,\omega_i) \qquad (4.22)$$

Since K is just a normalization constant, substituting (4.22) in (4.8) gives

$$f(B|\{X_s\}_1^{n-1} = \frac{\left[\sum_{i=1}^{K} P(V_{n-1}|B_i,\omega_i)P_i\right]}{\int \left[\text{numerator}\right] d\,B} f(B|\{X_s\}_1^{n-2}) \tag{4.23}$$

When the family of ith class, parameter-conditional c.d.f.'s has members continuous in X and B_i, such as the gaussian family, the decision equation is (4.6). $f(B|\{X_s\}_1^{n-1})$ is computed by (4.8). If the family is multivariate gaussian with X_s a single vector sample,

$$f(X_{n-1}|\omega_i,B_i) =$$

$$\frac{1}{(2\pi)^{\ell/2}|\Phi_{xx}^i|^{\frac{1}{2}}} \exp\left\{-\tfrac{1}{2}\,(X_{n-1}-\theta_i)^T \left[\Phi_{xx}^i\right]^{-1}(X_{n-1}-\theta_i)\right\} \tag{4.24}$$

where Φ_{xx}^i is the ith covariance matrix and θ_i the corresponding mean vector. Note that

$$B_i = (\Phi_{xx}^i,\ \theta_i)$$

$$B = (\{\Phi_{xx}^i\}_{i=1}^M,\ \{\theta_i\}_{i=1}^M,\ \{P_i\}_{i=1}^M)$$

The two types of optimum systems are shown in Fig. 2. The upper system uses the fixed bin model, and the lower system is for cases where the family has ith class, parameter-conditional c.d.f.'s continuous in X and B_i.

4.4 <u>Quantizing the Parameter Space</u>

The computation of $f(B|\{X_s\}_1^n)$ is iterative, in terms of $f(B|\{X_s\}_1^{n-1})$. The procedure is that, upon receiving sample X_n, $f(B|\{X_s\}_1^{n-1})$ is replaced in storage by $f(B|\{X_s\}_1^n)$. To store $f(B|\{X_s\}_1^{n-1})$, it is necessary that B take on a finite number of points in the parameter space. For some cases where there is supervision, it is not necessary to compute $f(B|\{X_s\}_1^{n-1})$. Instead, $f(X_n,\omega_i|\{X_s\}_1^{n-1})$ can be expressed in terms of a sufficient statistic

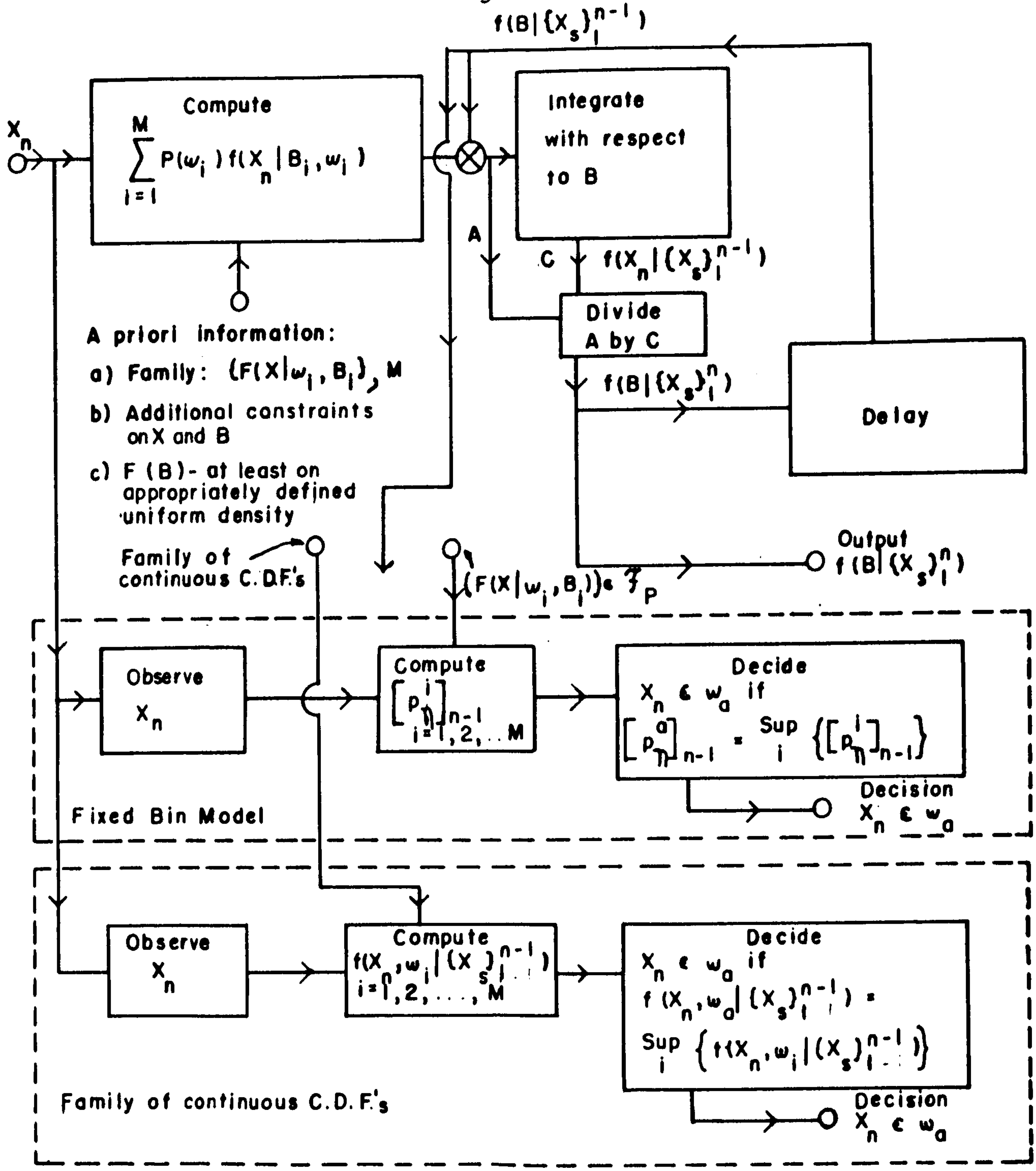

Fig. 2. Minimum probability of error systems

which is fixed in size[20].

In general, however, it is necessary to compute $f(B|\{X_s\}_1^{n-1})$. To do this, denote the number of scalar entries in B by q and write

$$B = (\phi_1, \phi_2, \ldots, \phi_q) \tag{4.25}$$

Quantize Q_i into N_i one dimensional levels of length Δ_i each, $i = 1,2,\ldots,q$. B can thus be in any of $\prod_{i=1}^{q} N_i$ q-dimensional levels. Denote a particular level by $L_{j_1,j_2,\ldots,j_q}$, and denote the true probability measure attached to this level by $m_{j_1,j_2,\ldots,j_q}$. Denote the probability measure attached to this level at the nth stage by $(m_{j_1,j_2,\ldots,j_q})_n$. Then, using this quantum level model,

$$f(B \in L_{j_1,j_2,\ldots,j_q}|\{X_s\}_1^n) = C(L)\ (m_{j_1,j_2,\ldots,j_q})_n \tag{4.26}$$

where C(L) is a normalization constant for the level considered. Equation (4.26) expresses the density of B in the level $L_{j_1,j_2,\ldots,j_q}$ at the nth stage in terms of the probability measure in that level at the (n-1)st stage.

Using the quantum model defined in this section, (4.7) can be written

$$(m_{j_1,j_2,\ldots,j_q})_n = \frac{f(X_{n-1}|L_{j_1,j_2,\ldots,j_q},\{X_s\}_1^{n-1})\ (m_{j_1,j_2,\ldots,j_q})_{n-1}}{\sum_{i=1}^{q}\sum_{j_i=1}^{N_i} f(X_{n-1}|L_{j_1,j_2,\ldots,j_q},\{X_s\}_1^{n-1})\ (m_{j_1,j_2,\ldots,j_2})_{n-1}} \quad \text{for all } j_i, q \tag{4.27}$$

If the samples are parameter-conditionally independent, (4.27) reduces to

$$(m_{j_1,j_2,\ldots,j_q})_n = \frac{f(X_{n-1}|L_{j_1,j_2,\ldots,j_q})\ (m_{j_1,j_2,\ldots,j_q})_{n-1}}{\sum_{i=1}^{q}\sum_{j_i=1}^{N_i} f(X_{n-1}|L_{j_1,j_2,\ldots,j_q})\ (m_{j_1,j_2,\ldots,j_q})_{n-1}} \tag{4.28}$$

$F(X_{n-1}|L_{j_1,j_2,\ldots,j_q})$ might be called a "level-conditional" mixture c.d.f. where the vector B characterizing the mixture has been quentized. It is a known function of the quantum levels.

As an example, let the family be multinomial parameter-conditional c.d.f.'s $\epsilon\ \mathcal{F}_{TP}$, and let there be sufficient constraints for identifiability. Consider a one dimensional (ℓ=1), binary (M=2) example such that

$$B = (p_1,\ p_2,\ldots,p_R,\ \theta_1,\ \theta_2,\ P_1) \tag{4.29}$$

Here B is an R+3 dimensional vector. B is then quantized and (4.28) applied, assuming the samples are parameter-conditionally independent.

As a second example, consider a one dimensional binary example where the family $\{F(X_s|\omega_i,B_i)\}$ is multivariate guassian with

$$\Phi^1_{xx} = \Phi^2_{xx} = \begin{bmatrix} \sigma^2 & o \\ o & \sigma^2 \end{bmatrix} \tag{4.30}$$

and σ_o, θ_{1_o}, and θ_{2_o} being fixed but unknown. Here,

$$B = (\theta_1,\ \theta_2,\ \sigma)$$

$$B_o = (\theta_{1_o},\ \theta_{2_o},\ \sigma_o),\ \text{the true vector} \tag{4.31}$$

and a sufficient constraint is

$$\theta_2 > \theta_1 \tag{4.32}$$

Computer simulated results for this last case were obtained, where the average sample-conditional error in making decisions on the nth sample was plotted vs. n in Fig. 3. If the number of experiments used to obtain this average is sufficiently large, then this average error vs. n is a computer simulation of the theoretical sample-conditional probability of error.

For $\sigma_o = 1$ and $P_{1_o} = 1/2$ and both known, and the constraint $\theta_2 > \theta_1$, and 90 quantum levels of length 1/10 in each dimension of the parameter space (8100 two-dimensional levels, 5040 having zero measure because of the constraint $\theta_2 > \theta_1$), and with θ_{1_o} and θ_{2_o} both unknown, the average error is plotted vs. n in Fig. 3 for the following 3 cases:

Case 1: $\theta_{1_o} = 0$, $\theta_{2_o} = 2.4$, and $F(\theta_1,\theta_2)$ uniform in the quantized parameter space.

Case 2: $\theta_{1_o} = 0$, $\theta_{2_o} = 0.5$, and $F(\theta_1,\theta_2)$ uniform in the qunatized parameter space.

Case 3: $\theta_{1_o} = -2$, $\theta_{2_o} = 2$, and $F(\theta_1,\theta_2) = (\frac{1}{2\pi})\exp(\frac{\theta_1-1}{2})^2 \exp(\frac{\theta_2-5}{2})^2$ in the qunatized parameter space.

For $P_{1_o} = 1/2$ and the constraint $\theta_2 > \theta_1$, σ_o, θ_{1_o}, and θ_{2_o} all unknown, with 45 quantum levels along the θ_1 and θ_2 axis and 10 along the σ axis, all of length 1/10, and with $F(\theta_1,\theta_2,\sigma)$ uniform in the quantized parameter space, the average error is plotted vs. $(\theta_{2_o} - \theta_{1_o})$ in Fig. 4 for two cases:

Case 1: n = 20, 10 experiments, $\theta_{1_o} = 0$, θ_{2_o} variable.

Case 2: n = 50, 10 experiments, $\theta_{1_o} = 0$, θ_{2_o} variable.

In this second example with three unknown, there were a total of 20,250 quantum levels, with zero measure in 10,570 levels because of the constraint $\theta_2 > \theta_1$.

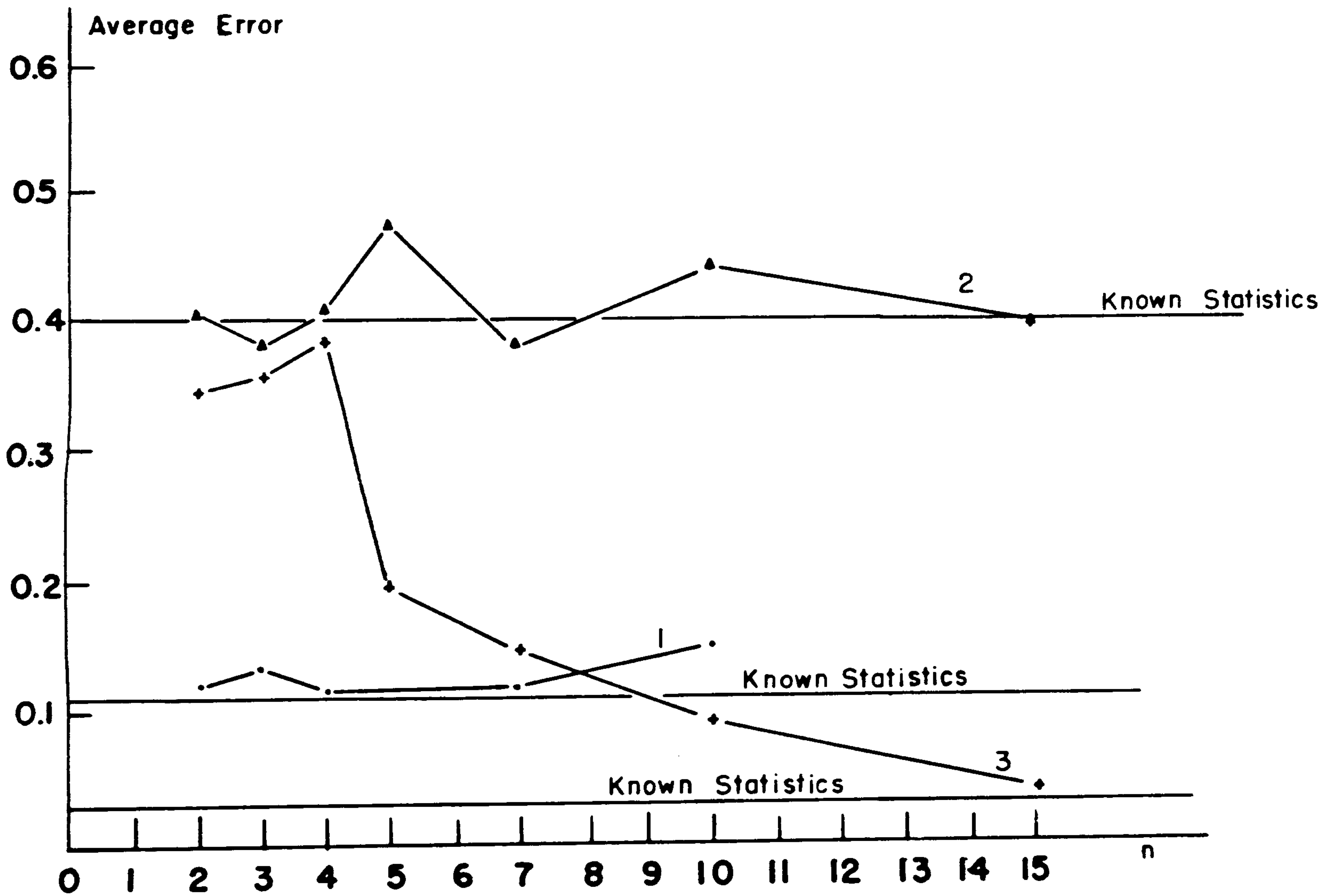

Fig. 3. Computer simulated average error vs n for binary gaussian example with two unknowns, θ_{1_0} and θ_{2_0}.

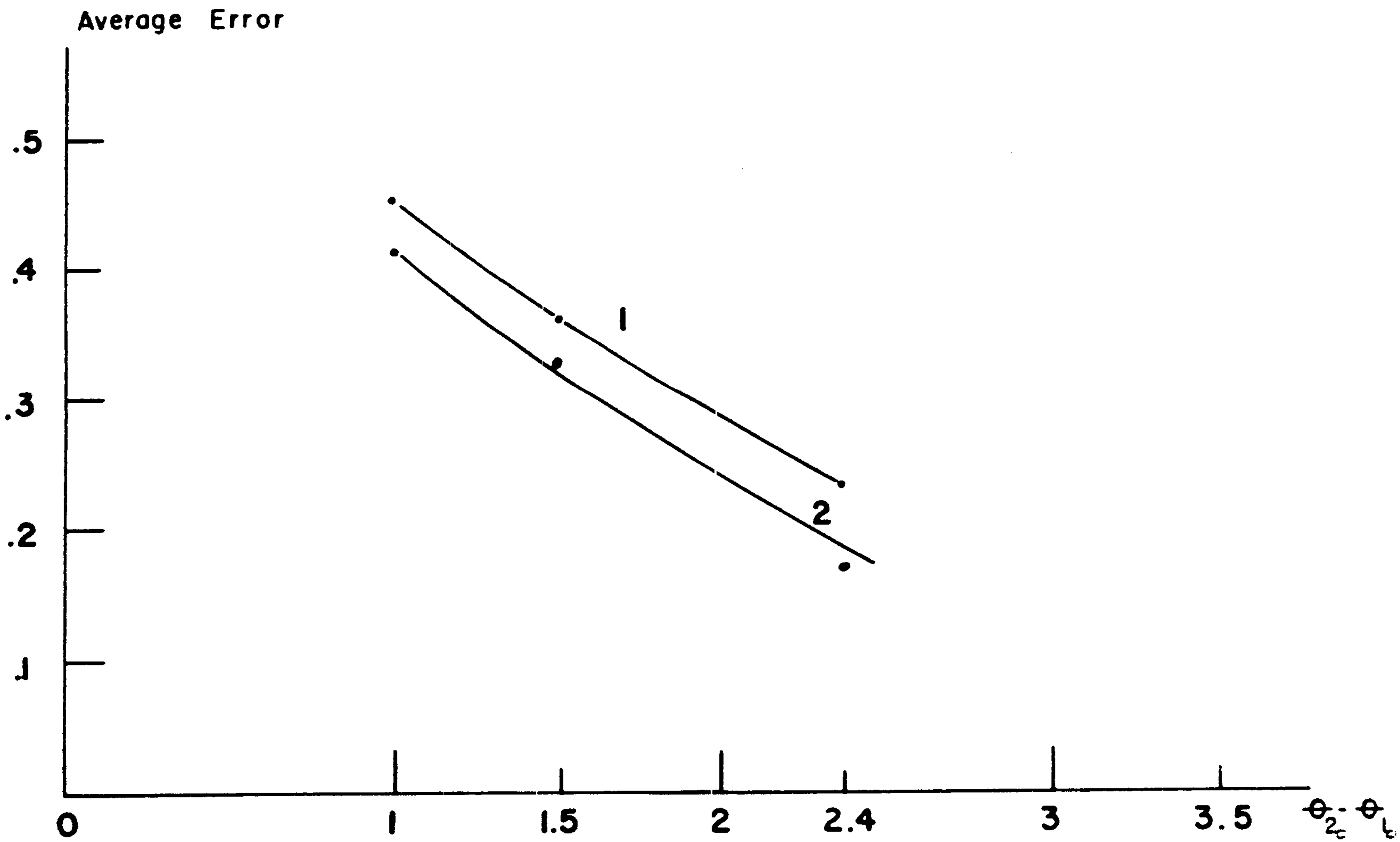

Fig. 4. Computer simulated average error vs $(\theta_{2_0}-\theta_{1_0})$ for binary gaussian example with three unknowns, θ_{1_0}, θ_{2_0} and P_{1_0}

CHAPTER V

CONSISTENT ESTIMATORS AND ASYMPTOTIC CONVERGENCE RATES

5.1 A Consistent Minimum Distance Estimator for B_o

Suboptimum systems which minimize probability of error as $n \to \infty$ but not sample-conditional probability of error can be designed using estimators. Since the family $\left\{F(X_s|\omega_i, B_i)\right\}$ is assumed known a priori, the decision equation (4.5) or (4.6) can be applied once the B_i are known. In Chapter III, it was shown that a consistent estimator for B_o can be found for the fixed bin model when the class of mixtures of multinomial distributions is identifiable. In this chapter we consider a consistent minimum distance estimator for B_o when $F(X|B)$ is continuous in X and B; we also obtain maximum likelihood estimators for the entries in B_o for a gaussian family of ith class, parameter-conditional c.d.f.'s. In addition, we obtain the asymptotic variance of these maximum likelihood estimators, equal with probability one, to that of the corresponding Bayes estimators.

Consider the nonsupervisory problem where $X = X_s = x_s$ is a one dimensional sample (ℓ=1) with a single class active on each sample and v=1.

Theorem 5.1

Let $x_{n_1}, \ldots, x_{n_n}$ be the order statistics with x_s, $s = 1,2,\ldots,n$, identically and independently distributed from the parameter-conditional mixture c.d.f. $F(x|B)$ continuous in x and B. Given the family $\{F(x|\pi_r, B_r)\}$, let the class of mixtures be identifiable. Then B_o can be estimated by a minimum distance estimator $\check{B}$ such that $\check{B} \xrightarrow{P} B_o$.

PROOF:

(i) Define $D_n = \sup_x |F_n(x) - F(x|B_o)|$, B_o being the true value of B

Then for any $\epsilon > 0$, Kolmogorov's Theorem[17,18] asserts that

$$\lim_{n\to\infty} P(D_n < \epsilon) = 1$$

That is, $F_n(x)$ converges in probability to $F(x|B_o)$ uniformly in x as $n \to \infty$, where B_o is the unique vector characterizing the mixture c.d.f. $F(x)$.

(ii) Obtain an estimator $\check{B}$ for B_o by solution (assuming the solution exists) for $\check{B}$ of

$$\inf_B \sup_x |F_n(x) - F(x|B)| = \sup_x |F_n(x) - F(x|\check{B})| \tag{5.1}$$

(iii) Since $\check{B}$ satisfies (5.1),

$$\sup_x |F_n(x) - F(x|\check{B})| \leq \sup_x |F_n(x) - F(x|B_o)|, \text{ all n.}$$

But since B_o is the true vector characterizing $F(x)$, we have for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(\sup_x |F_n(x) - F(x|\check{B})| \leq \sup_x |F_n(x) - F(x|B_o)| < \epsilon) = 1$$

or, for any ϵ,

$$\lim_{n\to\infty} P(\sup_x |F(x|\check{B}) - F(x|B_o)| < 2\epsilon) = 1$$

which gives $F(x|\check{B}) \xrightarrow{P} F(x|B_o)$. Since $F(X|B)$ is continuous in B, in particular at B_o, and there is a 1:1 mapping of $B \to F$ because of identifiability, this implies $\check{B} \xrightarrow{P} B_o$. This concludes the proof.

The fact that $\check{B} \xrightarrow{P} B_o$ implies that given the a priori knowledge required by Theorem 5.1, a system (not using all available a priori knowledge) can be designed which converges to the system obtainable had all statistics been known. Such a system is not sample-conditionally optimum but is, for all practical purposes, optimum in the limit $n \to \infty$. This method does not provide for taking into account all the a priori knowledge that the minimum conditional risk approach provided for.

5.2 Bayes and Maximum Likelihood Estimators for B_o

If $B = (\theta_1, \theta_2, \ldots, \theta_q)$ and a square law loss function $L(\hat{B}, B)$ is defined,

$$L(\hat{B}, B_o) = (\hat{\theta}_1 - \theta_{1_o})^2 + \ldots + (\hat{\theta}_q - \theta_{q_o})^2 \tag{5.2}$$

where

$$B_o = (\theta_{1_o}, \theta_{2_o}, \ldots, \theta_{q_o})$$

and

$$\hat{B} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q) \tag{5.3}$$

it is well known that a Bayes estimate B minimizing average loss is given by

$$\hat{B} = \int B \, f(B | \{X_s\}_1^n) \, d\,B \tag{5.4}$$

On the other hand, given the samples $\{x_s\}_1^n$ and the family of rth partition, parameter-conditional c.d.f.'s, the maximum likelihood estimator, $\tilde{B}$, for B_o is given by

$$f(\{x_s\}_1^n | \tilde{B}) = \sup_B f(\{x_s\}_1^n | B) \tag{5.5}$$

Theorem II of reference (20) gives rather general conditions under which $\tilde{B} \xrightarrow{P} B_o$ if and only if $\hat{B} \xrightarrow{P} B_o$. Since convergence in probability implies convergence in distribution, $F(\hat{B} | \{X_s\}_1^n) \to F(\tilde{B} | \{X_s\}_1^n)$ under these same conditions.

Thus, finding the asymptotic distribution of B also gives the asymptotic distribution of $\hat{B}$ in the sense that $|F(\hat{B} | \{X_s\}_1^n) - F(\tilde{B} | \{X_s\}_1^n)| \to 0$.

5.3 Implicit Equations for Maximum Likelihood Estimators

Consider the nonsupervisory problem where the family $\{F(X_s | \omega_i, B_i)\}$ is one dimensional gaussian. Then $B_i = (m_i, \sigma_i)$ and

$$f(x_s | \omega_i, B_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[-\frac{1}{2} \frac{(x_s - m_i)^2}{\sigma_i^2}\right] \tag{5.6}$$

If $M = 2$ (binary case) and $\sigma_i = \sigma$, $i = 1,2$,

$$B = (m_1, m_2, \sigma, P_1) \tag{5.7}$$

with the true value of B,

$$B_o = (m_{1_o}, m_{2_o}, \sigma_o, P_{1_o}) \tag{5.8}$$

It is shown in Appendix B that the maximum likelihood estimators, $\tilde{m}_1$, $\tilde{m}_2$, $\tilde{\sigma}$, and $\tilde{P}_1$ are given implicity as follows:

$$\tilde{P}_1 = \sum_{s=1}^{n} \phi_{1,s}, \quad \tilde{P}_2 = 1 - \tilde{P}_1 \tag{B-4}$$

$$\tilde{m}_i = \frac{\sum_{s=1}^{n} x_s \phi_{i,s}}{\sum_{s=1}^{n} \phi_{i,s}}, \quad i = 1,2 \tag{B-5}$$

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{s=1}^{n} \left[(x_s - \tilde{m}_1)^2 \phi_{1,s} + (x_s - \tilde{m}_2)^2 \phi_{2,s} \right] \tag{B-6}$$

where

$$\psi_s = \tilde{P}_1 \exp\left[-\frac{1}{2} \frac{(x_s - \tilde{m}_1)^2}{\tilde{\sigma}^2} \right] + (1 - \tilde{P}_1) \exp\left[-\frac{1}{2} \frac{(x_s - \tilde{m}_2)^2}{\tilde{\sigma}^2} \right] \tag{B-7}$$

$$\phi_{i,s} = \frac{\tilde{P}_i \exp\left[-\frac{1}{2} \frac{(x_s - \tilde{m}_i)^2}{\tilde{\sigma}^2} \right]}{\psi_s} \tag{B-8}$$

It can be shown that for the general case of M pattern classes where $\sigma_i = \sigma$, $i = 1,2,\ldots,M$, the maximum likelihood estimators are given implicity as follows:

$$\tilde{P}_i = \sum_{s=1}^{n} \phi_{i,s}, \quad i = 1,2,\ldots,M \tag{5.9}$$

$$\tilde{m}_i = \frac{\sum_{s=1}^{n} x_s \, \phi_{i,s}}{\sum_{s=1}^{n} \phi_{i,s}}, \quad i = 1,2,\ldots,M \tag{5.10}$$

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{s=1}^{n} \left[\sum_{i=1}^{M} (x_s - \tilde{m}_i)^2 \, \phi_{i,s} \right] \tag{5.11}$$

where

$$\psi_s = \sum_{i=1}^{M} \tilde{P}_i \exp\left[-\tfrac{1}{2} \frac{(x_s - \tilde{m}_i)^2}{\tilde{\sigma}^2} \right] \tag{5.12}$$

and $\phi_{i,s}$ is the same as for the binary case.

Returning to the binary case, even with $\hat{m}_1$, $\hat{m}_2$, and $\hat{\sigma}$ replaced by known values m_{1_o}, m_{2_o}, and σ_o, equation (B-4) for $\tilde{P}_1$ still involves $\tilde{P}_1$ on both sides of this equation:

$$\tilde{P}_1 = \sum_{s=1}^{n} \frac{\hat{P}_1 \exp\left[-\tfrac{1}{2} \frac{(x_s - m_{1_o})^2}{\sigma_o^2} \right]}{\hat{P}_1 \exp\left[-\tfrac{1}{2} \frac{(x_s - m_{1_o})^2}{\sigma_o^2} \right] + (1 - \tilde{P}_1) \exp\left[-\tfrac{1}{2} \frac{(x_s - m_{2_o})^2}{\sigma_o^2} \right]} \tag{5.13}$$

It is interesting to compare (5.13) with an estimator obtained by Robbins[12] for P_{1_o}. For the same situation, i.e. a gaussian family with known values m_{1_o}, m_{2_o}, and σ_o, Robbins' estimator is as follows:

$$\overset{x}{P}_1 = \frac{\sum_{s=1}^{n} \exp\left[-\tfrac{1}{2} \frac{(x_s - m_{1_o})^2}{\sigma_o^2} \right]}{\sum_{s=1}^{n} \exp\left[-\tfrac{1}{2} \frac{(x_s - m_{1_o})^2}{\sigma_o^2} \right] + \exp\left[-\tfrac{1}{2} \frac{(x_s - m_{2_o})^2}{\sigma_o^2} \right]} \tag{5.14}$$

Although (5.14) is an explicit solution for $\overset{x}{P}_1$ in terms of the samples, computer simulation shows the variance of $\overset{x}{P}_1$ is much larger than the variance of $\tilde{P}_1$. In a subsequent section, computer simulated results for the variance of $\widetilde{P}_1$ are presented as a function of n and other important parameters. The computer simulated results for the variance of $\overset{x}{P}_1$ will not be given since they give no indication of convergence.

5.4 Convergence and Asymptotic Distribution of $\widetilde{B}$

In this section we first state the sufficient conditions for maximum likelihood estimators to converge with a known asymptotic distribution. We then relate these sufficient conditions to the nonsupervisory problem and, in particular, derive the asymptotic distribution of the estimators given by (B-4) through (B-8).

Let X be an ℓ-dimensional vector with distribution $F(X|B_o)$, a parameter-conditional mixture with B_o given by (5.3). Define[17] the following:

$$T_j(X|B) = \frac{\partial}{\partial\theta_j} \text{Log } f(X|B) \tag{5.15}$$

$$T_{jK}(X|B) = \frac{\partial}{\partial\theta_K} T_j(X|B) \tag{5.16}$$

$$C_{jK}(B) = \int \left[T_j(X|B)\; T_K(X|B)\right] f(X|B)\; d\,X \tag{5.17}$$

$$D_{jK}(B) = \int \left[T_{jK}(X|B)\right] f(X|B)\; d\,X \tag{5.18}$$

It is said[17] that $F(X|B)$ <u>is regular with respect to its first θ_j - derivative</u> if

$$E\left[T_j(X|B)\right] = \frac{\partial}{\partial\theta_j} \int d\,F(X|B) = \int T_j(X|B)\; d\,F(X|B) = 0 \tag{5.19}$$

and $F(X|B)$ <u>is regular with respect to its second θ_j-derivative</u> if the matrix $\left[C_{jK}(B)\right]$ is positive definite and if

$$E\left[T_j(X|B)\; T_K(X|B)\right] + E\left[T_{jK}(X|B)\right] = \frac{\partial^2}{\partial\theta_j^2} \int d\,F(X|B) = 0 \tag{5.20}$$

We now state a theorem[17] giving sufficient conditions for $\widetilde{B}$ to converge a.c. to $B_o = (\theta_{1_o}, \theta_{2_o}, \ldots, \theta_{q_o})$.

Theorem 5.2. Let $\{X_s\}_1^n$ be n independent and identically distributed samples from the c.d.f. $F(X|B_o)$, where $B_o = (\theta_{1_o},\ldots,\theta_{q_o})$ and $F(X|B)$ is reqular with respect to its first θ_j derivatives. Let $T_j(X|B)$, $j = 1,2,\ldots,q$, be a continuous function of B for all values of X, except possibly a set of zero probability. Then there exists a sequence of solutions $\{(\tilde{\theta}_1,\ldots,\tilde{\theta}_q)\}$ which converge almost certainly to $(\theta_{1_o},\ldots,\theta_{q_o})$. If the solution is a unique vector $(\tilde{\theta}_1,\ldots,\tilde{\theta}_q)$ for $n \geq$ some n_o, the sequence of vectors converges almost certainly to $(\theta_{1_o},\ldots,\theta_{q_o})$ as $n \to \infty$.

Theorem 5.2 gives sufficient conditions for $\tilde{B}$ to converge a.c. to B_o. On the other hand, a necessary condition for there to be a unique solution for B_o, given $F(X)$ and $F(X|B)$, is that the class of mixtures be identifiable. This seems to imply that Theorem 5.2 gives a sufficient condition for identifiability. Actually, Theorem 5.2 assumes identifiability by the statement, "if the solution is a unique vector $(\tilde{\theta}_1,\ldots,\tilde{\theta}_q)$ for $n \geq$ some n_o ..."

If, in addition to satisfying the first regularity conditions, $F(X|B)$ satisfies the second regularity condition, then B is asymptotically normal according to the following theorem[17]:

Theorem 5.3: If $\{X_s\}_1^n$ is a sequence of independent and identically distributed samples from $F(X|B_o)$, where B_o has q entries and $F(X|B)$ is regular with respect to its first and second θ_j derivatives, and if $(\tilde{\theta}_1,\ldots,\tilde{\theta}_q)$ is unique for $n \geq$ some n_o, and measurable with respect to $\prod_{s=1}^{n} F(X_s|B)$, then $(\tilde{\theta}_1,\ldots,\tilde{\theta}_q)$ is asymptotically distributed for large n, according to the q-dimensional normal distribution $N(\{\theta_{j_o}\}_1^q, [n\, C_{jK}(B_o)]^{-1})$.

Returning to the binary gaussian nonsupervisory problem in Section 5.2, we can show that the requirements of Theorem 5.2 and Theorem 5.3 are fulfilled as follows:

a) The mixture is identifiable, according to Proposition A-1, if we constrain $m_{2_o} > m_{1_o}$, which would involve no loss of generality.

b) The first and second regularity conditions are shown to hold in Appendix C.

5.5 Theoretical and Computer Simulated Asymptotic Variances

Let asymptotic variances be the entries in $[n\ C_{jk}\ (B_o)]^{-1}$ corresponding to the asymptotic distribution of $(\tilde{\theta}_1,\ldots,\tilde{\theta}_q)$. Consider, for convenience, the case where there are two unknowns, $\theta_{1_o} = m_{1_o}$ and $\theta_{2_o} = m_{2_o}$. Performing the required matrix inversion and denoting the entries in $\left[n\ C_{jk}\ (B_o)\right]^{-1}$ by $E\left[\tilde{\theta}_j - \theta_{j_o}\right]^2$, we obtain

$$n\ E\left[\tilde{m}_i - m_{i_o}\right]^2 = \frac{C_{ii}(B_o)}{\left[C_{11}(B_o)C_{22}(B_o)-C_{21}(B_o)C_{12}(B_o)\right]} \tag{5.21}$$

where

$$C_{ii}(B_o) = \frac{P_{i_o}^2}{\sqrt{2\pi}\ \sigma_o^5} \int_{-\infty}^{\infty} \frac{(x-m_{i_o})^2\ e^{-\frac{(x-m_{i_o})^2}{\sigma_o^2}}}{P_{1_o}\ e^{-\frac{(x-m_{1_o})^2}{2\ \sigma_o^2}} + (1-P_{1_o})\ e^{-\frac{(x-m_{2_o})^2}{2\ \sigma_o^2}}}\ dx \tag{5.22}$$

Defining

$$H_i(x) = e^{-\frac{(x-m_{i_o})^2}{2\ \sigma_o^2}} \tag{5.23}$$

(5.22) becomes

$$C_{ii}(B_o) = \frac{P_{i_o}^2}{\sqrt{2\pi}\ \sigma_o^5} \int_{-\infty}^{\infty} \frac{(x-m_{i_o})^2\ [H_i(x)]^2}{P_{1_o} H_1(x) + (1-P_{1_o})\ H_2(x)}\ dx \tag{5.24}$$

Also

$$C_{ij}(B_o) = \frac{P_{i_o} P_{j_o}}{\sqrt{2\pi}\,\sigma_o^5} \int_{-\infty}^{\infty} \frac{(x-m_{i_o})(x-m_{j_o})\,H_i(x)\,H_j(x)}{P_{1_o} H_1(x) + (1-P_{1_o})\,H_2(x)}\,dx \tag{5.25}$$

Note that $C_{12} = C_{21}$ because of symmetry.

The necessary integrals (5.24) and (5.25) were evaluated, using a digital computer, and substituted into (5.21). The results, $n\,E[\tilde{m}_i - m_{i_o}]^2$, are plotted vs$(m_{2_o} - m_{1_o})$ in Fig. 5 for the following cases:

Case 1: $\sigma_o = 1$, $P_{1_o} = \frac{1}{2}$, both known; $m_{1_o} = 0$, m_{2_o} variable, both unknown.

Case 2: $\sigma_o = 0.5$, $P_{1_o} = \frac{1}{2}$, both known; $m_{1_o} = 0$, m_{2_o} variable, both unknown.

Case 3: $\sigma_o = 0.2$, $P_{1_o} = \frac{1}{2}$, both known; $m_{1_o} = 0$, m_{2_o} variable, both unknown.

The important aspects of the results in Fig. 5 are the following:

(a) $n[\tilde{m}_i - m_{i_o}]^2$ decreases as $(m_{2_o} - m_{1_o})$ increases

(b) $n[\tilde{m}_i - m_{i_o}]^2$ decreases as σ_o decreases

(c) A perturbation occurs in each curve. The curve corresponding to Case 1 with $\sigma_o = 1$ has the perturbation in the region where $f(x|B)$ changes from a bimodal to a unimodal density function. The value of $(m_{2_o} - m_{1_o})$ for this change decreases as σ_o decreases; this would explain the moving of the perturbation to the left in Fig. 5 as σ_o decreases.

(d) The value of $n\,E[\tilde{m}_i - m_{i_o}]^2$ increases indefinitely as $(m_{2_o} - m_{1_o})$ decreases to zero.

One way to explain (d) is as follows: it is assumed[17] in the proofs of Theorem 5.2 and Theorem 5.3 that the components of B_o are functionally independent so that inverse matrix $[C_{jk}]^{-1}$ exists. When $m_{2_o} = m_{1_o}$, this assumption is violated, and $[C_{jk}]^{-1}$ does not exist. If it is known a priori that $m_{1_o} = m_{2_o}$, then this problem is supervisory and the a priori assumption of two pattern classes in incorrect. Recall, however, that a priori knowledge of M is assumed in the approach taken in this thesis.

Consider next a special case of this binary, one dimensional, gaussian nonsupervisory problem, where only one of the four parameters m_{1_o}, m_{2_o}, σ_{1_o}, and P_{1_o}, is unknown. This is not a special case of Theorem 5.3 where all but one of the θ_{j_o}'s are known. Instead, the result is that

$$n\,E[\tilde{\theta}_j - \theta_{j_o}]^2 = \frac{1}{C_{jj}(\theta_{j_o})},\ \text{only } \theta_{j_o} \text{ unknown} \tag{5.26}$$

Equation (5.26) for $\theta_j = P_1$ was evaluated using a digital computer. The results, $n\,E[\tilde{P}_1 - P_{1_o}]^2$ vs$(m_{2_o} - m_{1_o})$, are plotted in Fig. 6 for the following cases (with $m_{1_o} = 0$, m_{2_o} variable, $\sigma_{1_o} = 1$, all known):

Case 1: $P_{1_o} = 0.5$, unknown.

Case 2: $P_{1_o} = 0.66$, unknown.

Case 3: $P_{1_o} = 0.75$, unknown.

Then, for three values of $(m_{2_o} - m_{1_o})$, $E[\tilde{P}_1 - P_{1_o}]^2$ is plotted vs n in Fig. 7, using the results displayed in Fig. 6.

To check the theoretical results given in Fig. 7, the quantity Av. $(\tilde{P}_1 - P_{1_o})^2$ was simulated using a digital computer, by evaluating (B-4) as a function of n. Given the samples $\{X_s\}_1^n$, an iterative solution of (B-4) was obtained; the iteration was started for the first sample by choosing $\tilde{P}_1$ on the right side of (B-4) from a uniform [0,1] random number generator. To obtain the average, 100 experiments were performed for each value of n considered. These computer simulated results are presented in Fig. 8 for comparison with the corresponding theoretical results of Fig. 7. The essential conclusion is that agreement between computer simulation and theory improves as $(m_{2_o} - m_{1_o})$ increases and as n increases. The latter is certainly to be expected since the theoretical curves in Fig. 6 are asymptotic results.

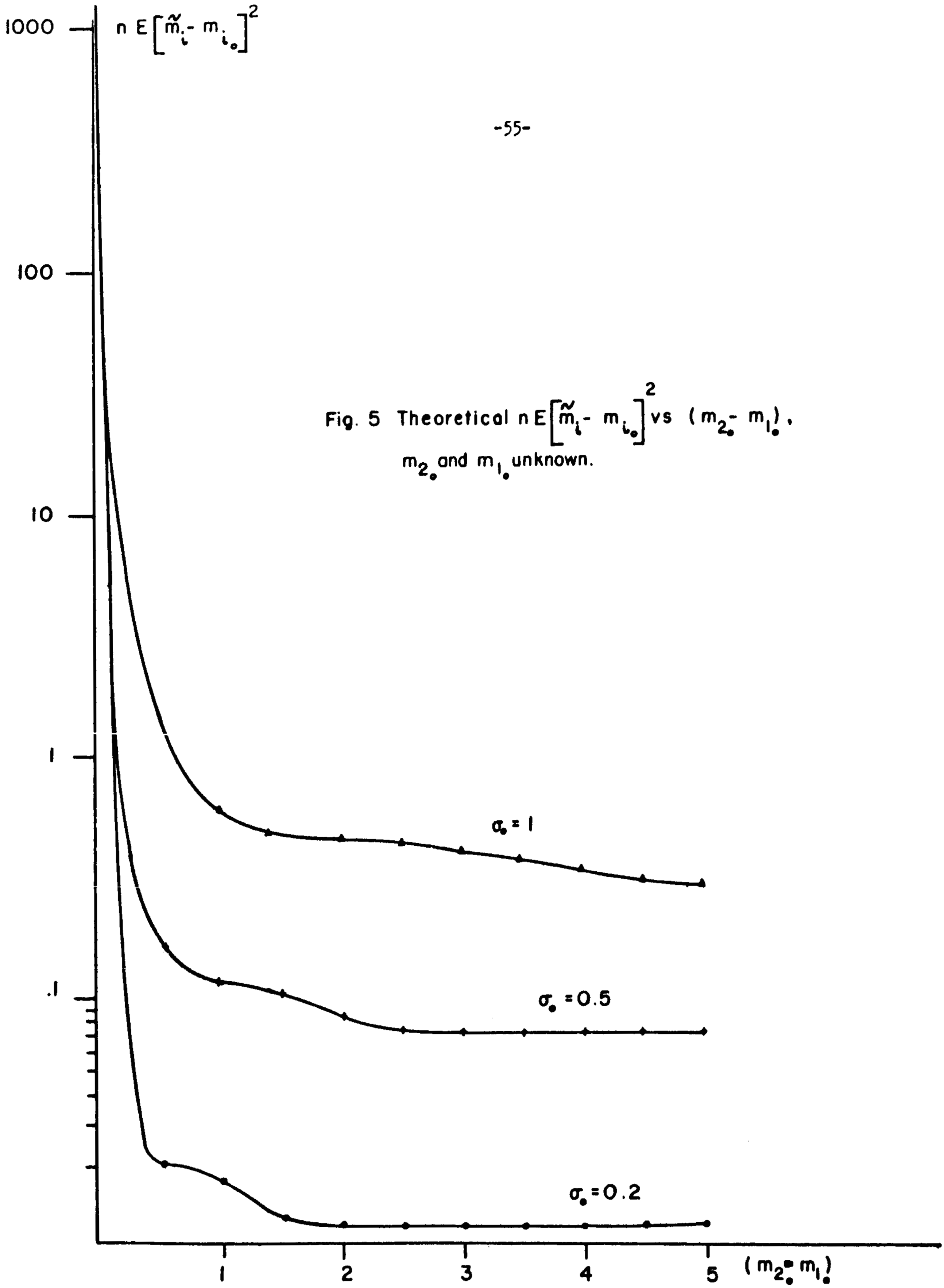

Fig. 5 Theoretical $n\,E\left[\tilde{m}_i - m_{i_0}\right]^2$ vs $(m_{2_0} - m_{1_0})$, m_{2_0} and m_{1_0} unknown.

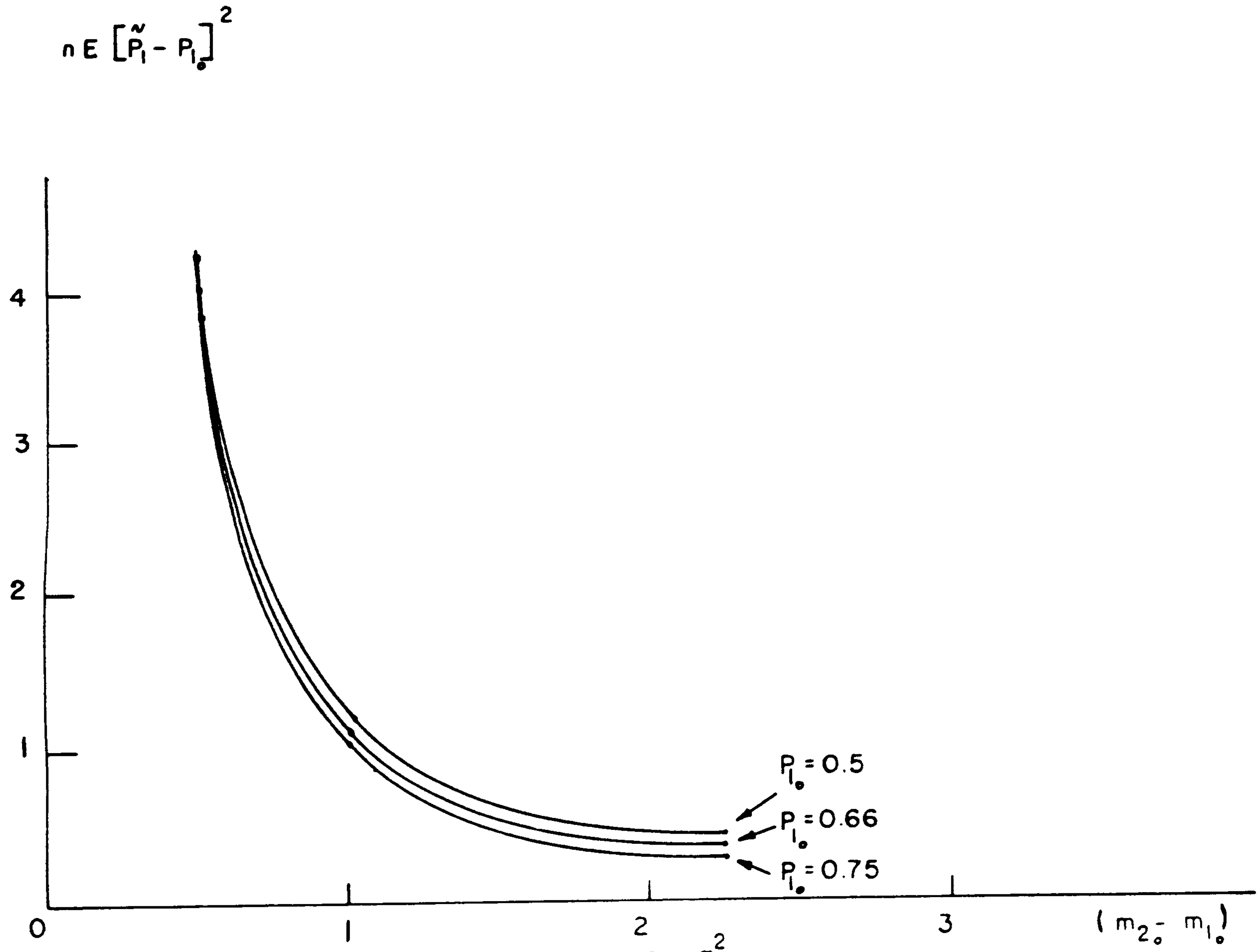

Fig. 6 Theoretical $n\,E\left[\tilde{P}_1 - P_{1_0}\right]^2$ vs. $(m_{2_0} - m_{1_0})$ for P_{1_0} unknown.

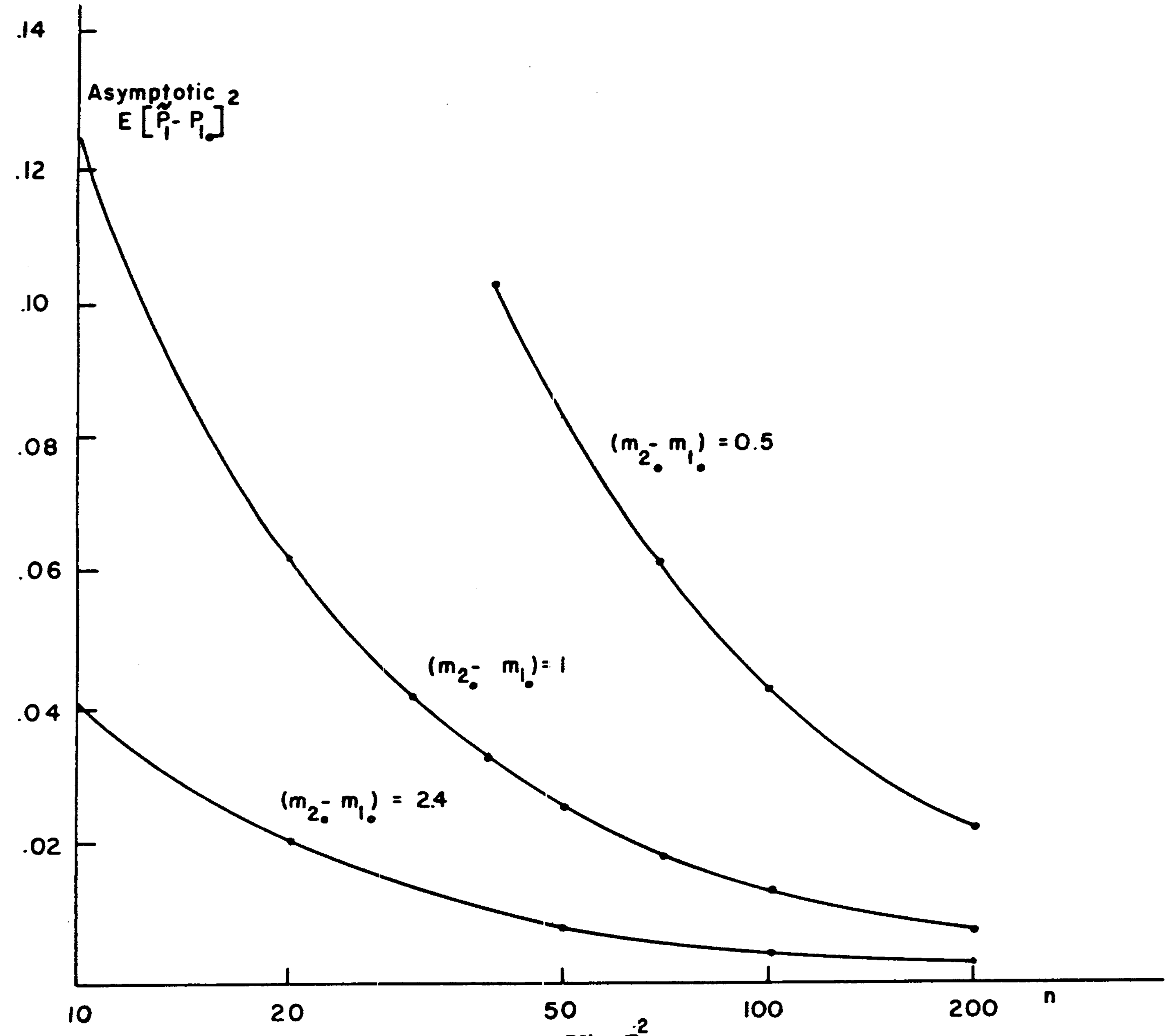

Fig. 7 Theoretical $E[\tilde{P}_1 - P_{1_0}]^2$ vs. n for P_{1_0} unknown.

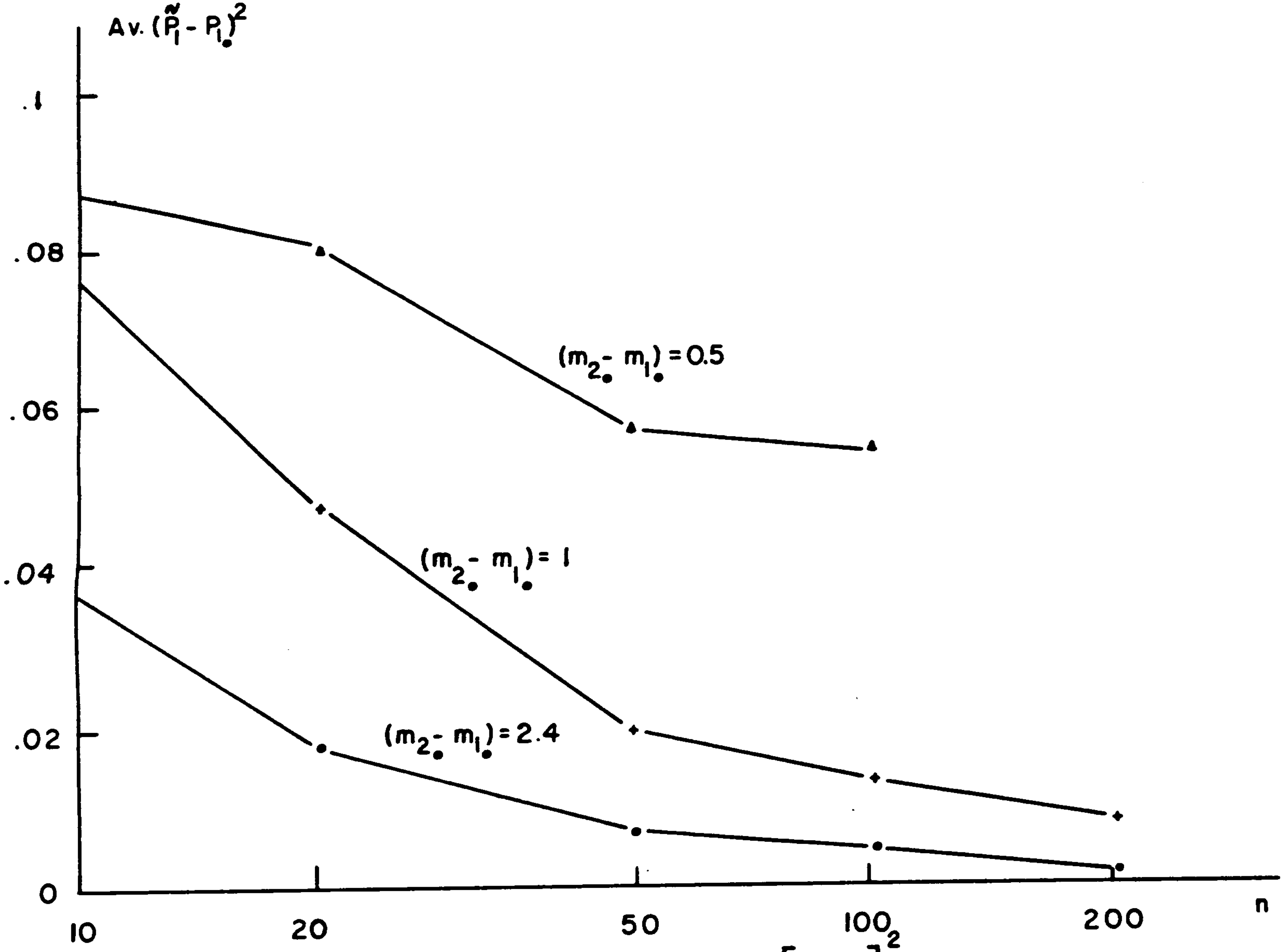

Fig. 8 Computer simulated av. $\left[\tilde{P}_1 - P_{1_\bullet}\right]^2$ vs. n for $P_{1_\bullet}$ unknown.

CHAPTER VI

CONCLUSIONS

6.1 General Conclusions

Nonsupervisory problems lack the a priori knowledge of sample classification. For this reason, the probability distribution function for the samples is in general more complex than when there is supervision. There are nonsupervisory problems where the distribution function (mixture c.d.f.) for the samples is not uniquely characterized by the mixing parameters and the parameters characterizing each ith class-conditional c.d.f. It is not possible to estimate these parameters with consistent estimators or to optimally converge to a minimum-probability-of-error solution. By providing additional a priori knowledge about the ith class-conditional c.d.f.'s, the way the samples are taken, spacial constraints, constraints on the parameters characterizing the mixture c.d.f., etc., the nonsupervisory problem may have a solution. Even when the ith class-conditional c.d.f.'s are one dimensional gaussian, we are not assured of a solution without sufficient constraints on the parameters. These constraints cause no loss of generality in this gaussian case, but must be imposed.

The importance of sufficient a priori knowledge in nonsupervisory problems is exemplified when the rth-partition, parameter-conditional c.d.f.'s are empirical c.d.f.'s, corresponding to no a priori knowledge about the c.d.f.'s. Here the resulting c.d.f. of the samples is characterized by the ordered samples, the number of which increases as n increases. A parameter-conditional mixture does not exist for this problem. If the ith-class conditional c.d.f.'s are approximated by multinomial c.d.f.'s, the number of parameters characterizing the mixture is fixed in size; but it is not possible to estimate these parameters in general without additional a priori knowledge. On the other hand, when the samples are classified, estimating parameters characterizing such ith class-multinomial c.d.f.'s corresponds to the histogram concept. It can be concluded that such nonparametric techniques do not directly apply to nonsupervisory problems. The difficulty is that such nonparametric techniques do not directly provide for the use of additional a priori knowledge. By taking into account additional a priori knowledge, such as that mentioned above, a nonsupervisory problem may have a solution. We have introduced a construction technique, where additional a priori knowledge, such as spacial constraints, symmetry, the number of samples taken at the sth. observation, etc., is utilized.

6.2 Conclusions on Performance

Evaluation of the theoretical performance of the optimum systems is in general difficult. One approach, given in Sections 5.2, 5.3, 5.4, and 5.5, is to find the asymptotic distribution of Bayes estimators or maximum likelihood estimators for the parameters characterizing the mixture concerned. Using classical statistical techniques, it was shown that the joint distribution of these estimators is multivariate gaussian (when a solution exists). Using this joint distribution, a bound on asymptotic sample-conditional probability of error can be obtained.

An example where the above asymptotic distribution is evaluated was given in Section 5.5 for the binary, gaussian case with two unknowns. Theoretically, the asymptotic distribution for the M-ary gaussian case with any number of unknowns can be obtained. Practically, however, this requires evaluating a large number of integrals using a digital computer.

Results giving tight bounds on sample-conditional probability of error for all values of n would be useful. Such results, however, should be presented with a precise statement of the a priori knowledge utilized in the nonsupervisory problem concerned.

6.3 Implementation Difficulties

The general optimum system, illustrated in Fig. 2, can be implemented using a digital computer by quantizing the parameter space B (c.f Section 4.4). Once the parameter space is quantized, the required storage in a digital computer is fixed in size--it does not grow as n increases. On the other hand, this fixed amount of storage increases as the number of unknowns in B increases. As this storage increases, the number of computations a digital computer must perform, in the time interval between receiving two samples, increases. The speed with which the computer operates can be held constant, however, if the time interval between samples is increased.

A computer simulation of an optimum nonsupervisory problems having three unknowns was given in Section 4.4. It is difficult in general to implement such a problem when it has more than four unknowns without having more storage than that available in an IBM 7094 computer.

There are some specific nonsupervisory problems with certain spacial a priori knowledge (Section 3.4) which have many unknowns, say 100, that can be implemented with an IBM 7094 computer. This reemphasizes the need for precisely stating the a priori knowledge assumed used in a nonsupervisory problem.

Besides digital implementation, it is possible to use analog techniques to implement optimum systems. For example, assume that it takes T/2 seconds to obtain sample X_s; and sample X_{s+1} begins to be received T/2 seconds after

X_s is completely received. In the T/2 seconds between samples, all computations in the iteration (4.23) must be made. As the number of unknowns in the nonsupervisory problem increases, the computation rate between samples increases. This computation rate determines the bandwidth required in a delay line used for storage in the analog implementation. We can thus conclude that for a given sample transmission rate, the bandwidth required in the analog processing equipment increases as the number of unknowns in the nonsupervisory problem increases.

In summary, digital implementation of optimum nonsupervisory problems is restricted by increasing required storage as the number of unknowns increases. And, analog implementation of optimum nonsupervisory problems is restricted by increasing required delay-line bandwidth as the number of unknowns increases.

BIBLIOGRAPHY

1. Abramson, N., and D. Braverman, "Learning to Recognize Patterns in a Random Environment," IRE International Symposium on Information Theory, Vol. IT-8, pp. 558-563, July, 1962.

2. Fralick, S.C., "The Synthesis of Machines Which Learn Without a Teacher," Stanford Tech. Report No. 61308-9, April, 1964.

3. Daly, R.F., "The Adaptive Binary-Detection Problem on the Real Line," Stanford Electronics Laboratory Report TR 2003-3, February, 1962.

4. Cooper, D.B., and P.W. Cooper, "Nonsupervised Adaptive Signal Detection and Pattern Recognition," Information and Control, Vol. 7, No. 3, September, 1964.

5. Sebestyen, G.S., "Pattern Recognition by an Adaptive Process of Sample Set Construction," IRE International Symposium on Information Theory, Vol. IT-8, pp. 582-591, July, 1962.

6. First Semi-Annual Research Summary, School of Electrical Engineering, Purdue University, July through December, 1964.

7. Patrick, E.A., and J.C. Hancock, "The Nonsupervised Learning of Probability Spaces and Recognition of Patterns," IEEE International Convention, Information II, March, 1965.

8. Sebestyen, G.S., Decision-Making Processes in Pattern Recognition, Macmillan Company, New York, 1962.

9. Teicher, Henry, "On the Mixture of Distributions," Annals of Math. Stat., Vol. 31, pp. 55-73, 1960.

10. Teicher, Henry, "Identifiability of Finite Mixtures," Annals of Math. Stat., Vol. 34, No. 4, December, 1963.

11. Blischke, W.R., "Moment Estimators for the Parameters of Two Binomial Distributions," Annals of Math. Stat., Vol. 33, pp. 444-454, 1962.

12. Robbins, Herbert, "The Empirical Bayes Approach to Statistical Decisions Problems," Annals of Math. Stat., Vol. 35, pp. 1-20, 1964.

13. Fraser, D.A.S., Nonparametric Methods in Statistics, John Wiley and Sons, New York, 1957.

14. Fralick, S.C., "Learning to Recognize Patterns Without a Teacher," Stanford Technical Report 6103-10, SEL-65-011, March 1965.

15. Keehn, Daniel G., "A Note on Learning for Gaussian Properties," PGIT, pp. 126-132, January, 1965.

16. Second Semi-Annual Research Summary, School of Electrical Engineering, Purdue University, January through June, 1965.

17. Wilks, Samuel S., Mathematical Statistics, John Wiley and Sons, New York, 1962.

18. Fisz, Probability Theory and Mathematical Statistics, John Wiley and Sons, New York, 1963.

19. Hancock, J.C., and E.A. Patrick, "Iterative Computation of A Posteriori Probability for M-ARY Nonsupervised Adaptation," submitted to technical journal June 1, 1965.

20. Spragins, J.D., "Reproducing Distributions for Machine Learning," Stanford Electronics Laboratories Technical Report No. 6103-7, November, 1963.

21. Mood and Graybill, Introduction to the Theory of Statistics, McGraw-Hill Book Co., New York, pp. 192, 1963.

APPENDIX A

MIXTURES AND IDENTIFIABILITY

Following Teicher's definition[10] of identifiability for one dimensional mixture c.d.f.'s, we give the following definition of identifiability for ℓ-dimensional mixture c.d.f.'s.

Identifiability of Mixture C.D.F.'s

Let $\mathcal{F} = \{F(X|\alpha) : \alpha \in R_1^k\}$ constitute a family of ℓ-dimensional index-conditional c.d.f.'s, indexed by a point α in a subset R_1^k of Euclidean k-space R^k. Then, the ℓ-dimensional mixture c.d.f.

$$F(X) = \int_{R_1^k} F(X|\alpha) \, d\, G(\alpha) \tag{A-1}$$

is the image under the above mapping, say $\tilde{\mathcal{F}}$, of the k-dimensional c.d.f. G(where the measure μ_G induced by G assigns measure one to R_1^k).

The c.d.f. F(X) is called a mixture (or G-mixture of $\mathcal{F}$) while G is referred to as the mixing c.d.f. Let $\mathcal{G}$ denote the class of all such c.d.f.'s G, and $\mathcal{H}$ the induced class of mixtures F(X) (given a priori the family $\mathcal{F}$). Then $\mathcal{H}$ will be said to identifiable if $\tilde{\mathcal{F}}$ is a one-to-one map of $\mathcal{G}$ onto $\mathcal{H}$.

F(X) is called a finite mixture if its mixing distribution G, or rather the corresponding measure μ_G, is discrete and doles out positive mass to only a finite number (W) of partitions in R_1^k. Let these partitions be π_r, $r = 1,2,\ldots,W$, and the corresponding mass or measure be $P(\pi_r)$, $r = 1,2,\ldots,W$. Then (A-1) becomes

$$F(X) = \sum_{r=1}^{W} F(X|\pi_r)\, P(\pi_r) \tag{A-2}$$

Identifiability of Parameter Conditional Mixture C.D.F.'s

Let $\mathcal{F} = \{F(X|\pi_r, B_r) : \pi_r \in R_1^k,\}_{r=1}^{W}$ constitute a family of size W of

ℓ-dimensional rth partition, parameter-conditional c.d.f.'s, the rth partition indexed by π_r; and let B_r be the vector set of parameters characterizing the rth partition. Then the ℓ-dimensional, parameter-conditional mixture c.d.f.

$$F(X|B) = \sum_{r=1}^{W} F(X|B_r,\pi_r)\ P(\pi_r)$$

is the image under the above mapping, say $\tilde{\mathcal{F}}$, of the vector parameters $B_1,B_2,\ldots,B_W,\ P(\pi_1),\ldots,P(\pi_W)$, where $B = (B_1,B_2,\ldots,B_W,\ \{P(\pi_r)\}_1^W)$.

Let G denote the class of all such sets of mixing parameters $\{P(\pi_r)\}_1^W$ and vector parameters $\{B_r\}_1^W$, and $\mathcal{H}$ the induced class of parameter-conditional mixtures $F(X|B)$ (given a priori the family $\mathcal{F}$). Then $\mathcal{H}$ will be said to be <u>identifiable</u> if $\tilde{\mathcal{F}}$ <u>is a one-to-one map of B onto $\mathcal{H}$</u>.

Thus, for a given c.d.f. $F(X)$, there is a unique vector B_o such that $F(X) = F(X|B_o)$.

The following is a simple extension of Teicher's Theorem on identifiability to the case of parameter-conditional mixtures.

<u>Theorem A.1</u>. Let $\mathcal{F} = \{F(X|\pi_r,B_r)\}$ be a family of rth partition parameter-conditional c.d.f.'s with transforms $\phi_r(v_1,\ldots,v_\ell|B_r)$ defined for $V = (v_1,\ldots,v_\ell)$ and S_{ϕ_r} (the domain of definition of ϕ_r), such that the mapping $A:F \to \phi$ is linear and one-to-one. Suppose that there exists a total ordering ($\preceq$) of $\mathcal{F}$ such that $F_1 < F_2$ implies (i) $S_{\phi_1} \subseteq S_{\phi_2}$, (ii) the existence of some $V_1 \in \bar{S}_{\phi_1}$ (V_1 being independent of ϕ_2) such that $\lim_{V \to V_1} \dfrac{\phi_2(V)}{\phi_1(V)} = 0$. Then the class $\mathcal{H}'$ of all finite parameter-conditional mixtures of $\mathcal{F}$ is identifiable.

PROOF:

Suppose there are two finite sets of elements of $\mathcal{F}$, say $\mathcal{F}_1 = \{F_i,\ 1 \leq i \leq k\}$ and $\mathcal{F}_2 = \{\hat{F}_j,\ 1 \leq j \leq \hat{k}\}$, $F_i = F(X|\pi_i,\ B_i)$, $\hat{F}_j = F(X|\pi_j,\hat{B}_j)$, such that

$$\text{(a)} \quad \sum_{i=1}^{k} C_i F_i(X) \underset{x}{\equiv} \sum_{j=1}^{\hat{k}} \hat{C}_j \hat{F}_j(X) \quad 0 < C_i, \hat{C}_j \leq 1,$$

$$\sum_{i=1}^{k} C_i = \sum_{j=1}^{k} \hat{C}_j = 1$$

Without loss of generality, index the c.d.f.'s so that $F_i < F_j$, $\hat{F}_i < \hat{F}_j$, for $i < j$. If $F_1 \neq \hat{F}_1$, suppose also without loss of generality that $F_1 < \hat{F}_1$. Then, $F_1 < \hat{F}_j$, $1 \leq j \leq \hat{k}$ and from the transform (ed) version of (a), it follows that for $V \in V_1 = S_{\not{0}_1}: \left[V: \not{0}_1(V) \neq 0\right]$,

$$C_1 + \sum_{i=2}^{k} C_i\left[\not{0}_i(V)|\not{0}_1(V)\right] \underset{V}{\equiv} \sum_{j=1}^{\hat{k}} \hat{C}_j \left[\hat{\not{0}}_j(V)|\not{0}_1(V)\right]$$

Letting $V \to V_1$ through values in V_1, $C_1 = 0$ contradicting the supposition of (a) that $C_1 > 0$. Thus, $F_1 = \hat{F}_1$ and for any $V \in V_1$

$$(C_1 - \hat{C}_1) + \sum_{i=2}^{k} C_i\left[\not{0}_i(V)|\not{0}_1(V)\right] \underset{V}{\equiv} \sum_{j=2}^{k} \hat{C}_j \left[\hat{\not{0}}_j(V)|\not{0}_1(V)\right]$$

Again letting $V \to V_1$ through values in V_1, $C_1 = \hat{C}_1$ whence

$$\sum_{i=2}^{k} C_i F_i(X) \underset{x}{\equiv} \sum_{j=2}^{\hat{k}} \hat{C}_j \hat{F}_j(X)$$

Repeating the priori argument a finite number of times, we conclude that $F_i = \hat{F}_i$ and $C_i = \hat{C}_i$ for $i = 1,2,\ldots, \min(k,\hat{k})$. Further, if $k \neq \hat{k}$, say $k > \hat{k}$, then $\sum_{i=\hat{k}+1}^{k} C_i F_i(X) = 0$ implying $C_i - 0$, $\hat{k} + 1 \leq i \leq k$ in contradiction to (a). Thus, $k = \hat{k}$, $C_i = \hat{C}_i$ and $F_i = \hat{F}_i$ $1 \leq i \leq k$, implying $\mathcal{F}_1 = \mathcal{F}_2$ and identifiability of $\mathcal{H}$. That is, $B = \hat{B}$.

Proposition A.1. The class of one dimensional parameter-conditional mixtures of rth partition, parameter-conditional normal c.d.f.'s, with constraint that the family be ordered lexicographically by $N_i < N_j$ if $\sigma_i > \sigma_j$ or if $\sigma_i = \sigma_j$ but $\theta_i < \theta_j$, is identifiable.

PROOF:

Let $N_r = F(x|\theta_r, \sigma_r, \pi_r)$ denote the rth partition, parameter-conditional normal c.d.f. with mean θ_r and variance $\sigma_r^2 > 0$. Its bilateral Laplace transform is given by $\theta_r(v|\theta_r, \sigma_r^2) = \exp\{\sigma_r^2 t^2|2 - \theta_r t\}$. Order the family lexicographically by $N_i < N_j$ if $\sigma_i > \sigma_j$ or if $\sigma_i = \sigma_j$ but $\theta_i < \theta_j$. Then Theorem A.1 applies with $S_\emptyset = (-\infty, \infty)$ and $V_1 = v_1 = +\infty$

The significance of Proposition A.1 is that if the family of rth partition, parameter-conditional c.d.f.'s is one dimensional gaussian, then, given $F(X|B)$, there is a unique solution for $B_1 \cdots B_M, B_{M+1}$ if the a priori knowledge includes

(a) $\sigma_i > \sigma_j$, $i < j$ or

(b) if k is the smallest index such that $\sigma_k = \sigma_{k+1}$, then $m_k < m_{k+1}$

(c) repeat (a) and (b) starting with $\sigma_{k+1} > \sigma_{k+2}$, etc.

In other words, (a) ... (c) is sufficient a priori knowledge to assure identifiability. It is not necessary a priori knowledge to assure identifiability. We can veiw (a) ... (c) as a <u>constraint on the domain of definition of B</u>. If this constraint is utilized, then a unique solution for B_o can be found given the sequence of samples $\{x_s\}_1^n$ as $n \to \infty$.

The following is a proposition where we have simply extended Proposition A.1 to the multidimensional case.

<u>Proposition A.1'</u>: The class of mixtures of two (M = 2) one dimensional parameter conditional normal c.d.f.'s, $F(x|\omega_1,\theta_1,\sigma)$, $F(x|\omega_2,\theta_2,\sigma)$, with σ, θ_1, and P_1 known, is identifiable.

PROOF:

Suppose there are two finite sets of elements of $\mathcal{F}$, say $\mathcal{F}_1 = \{F_i\}_{i=1}^2$ and $\mathcal{F}_2 = \{\hat{F}_j\}_{j=1}^2$, $F_i = F(x|\theta_i,\sigma,\omega_i)$ $\hat{F}_j = F(x|\theta_j,\sigma,\omega_j)$, such that

$$\text{(a)} \quad \sum_{i=1}^{2} C_i F_i(x) \underset{x}{\equiv} \sum_{j=1}^{2} \hat{C}_j \hat{F}_j(x), \qquad 0 < C_i, \hat{C}_j < 1,$$

$$\sum_{i=1}^{2} C_i = \sum_{j=1}^{2} C_j = 1$$

Taking the bilateral Laplace transform of both sides of (a), we obtain

$$C_1 e^{\frac{\sigma^2 t^2}{2} - \theta_1 t} + C_2 e^{\frac{\sigma^2 t^2}{2} - \theta_2 t} \underset{t}{\equiv} \hat{C}_1 e^{\frac{\sigma^2 t^2}{2} - \theta_1 t} + \hat{C}_2 e^{\frac{\sigma^2 t^2}{2} - \hat{\theta}_2 t}$$

Since P_1 is known, $C_1 = \hat{C}_1$

$$C_2 = \hat{C}_2$$

$$\therefore e^{\frac{\sigma^2 t^2}{2} - \theta_2 t} \underset{t}{\equiv} e^{\frac{\sigma^2 t^2}{2} - \hat{\theta}_2 t}$$

$$\text{i.e.} \quad e^{(\hat{\theta}_2 - \theta_2)t} \underset{t}{\equiv} 1, \ \hat{\theta}_2 = \theta_2.$$

<u>Proposition A.2</u>. Let $\{F(X|\pi_r, B_r)\}$ be a finite family of ℓ-dimensional normal c.d.f.'s with $B_r = (M_r, \Phi^r_{xx})$ with mean vector $M_r = (m_{r_1}, m_{r_2}, \ldots, m_{r_\ell})$ and covariance matrix $\Psi^r_{xx} = [\sigma^r_{jk}]$. If the family is ordered lexicographically so that $N_1 < N_2 < N_3 <, \ldots, < N_W$ if $\sigma^1_{11} > \sigma^2_{11}, \ldots, \sigma^k_{kk} > \sigma^{k+1}_{kk}, \ldots$, or if $\sigma^k_{kk} = \sigma^{k+1}_{kk}$ but $m_{kk} < m_{k+1,k}$, then the family is identifiable.

PROOF:

The bilateral Laplace transform of $F(X|\pi_r, B_r)$ is given by $\phi_r = \exp(\frac{1}{2} V^T \Phi^r_{xx} V - M^T_r V)$. Then, with the family ordered as above, Theorem A.1 applies with $S_{\phi_r} = (-\infty < v_r < \infty,\ r = 1,2,\ldots,\ell)$ and $V_1 = (v_1 = \infty,\ v_r$ finite, $r = 2,3,\ldots,\ell)$.

Proposition A.3. The class of all finite mixtures of ℓ-dimensional c.d.f.'s which differ only by translational vectors and have bilateral Laplace transforms is identifiable if the family is ordered lexicographically by $F_1 < F_2$ if $m_{11} < m_{21}$.

PROOF:

Let $F(X|\pi_r, B_r) = F_o(X - M_r)$ denote the rth partition, parameter-conditional c.d.f. with $B_r = M_r$, a mean vector, which differs from other partitioned parameter-conditional c.d.f.'s only by M_r. If $V = (j\gamma_1 + \alpha_1, \ldots, j\gamma_\ell + \alpha_\ell) \in S_\emptyset$ and $\emptyset_o(V)$ is the bilateral Laplace transform of $F_o(X)$, then $\exp(-V^T M_r)\ \emptyset_o(V)$ is the bilateral Laplace Transform of $F(X|\pi_r, B_r)$. Order the family lexicographically by $F_1 < F_2$ if $m_{11} < m_{21}$. Then Theorem A.1 applies with $S_\emptyset =$ ℓ-dimensional complex Euclidean vector space and

$$V_1 = (j0 + \infty, j\gamma_2 + \alpha_2, \ldots, j\gamma_\ell + \alpha_\ell;\ \gamma_i, \alpha_i \text{ finite, } i = 2, \ldots, \ell)$$

We have been concerned with a vector X and its c.d.f. and have not related X to a specific nonsupervisory problem. Thus the results so far on mixtures and identifiability are quite general. They can be applied to the several nonsupervisory problems defined in Chapter II and the families of ith class, parameter-conditional c.d.f.'s defined in Chapter III.

The families of the ith class, parameter-conditional c.d.f.'s defined in Chapter III have members which are multinomial distributions. These multinomial distributions arise when a general family of ith class-conditional c.d.f.'s are approximated by ith class-conditional multinomial distributions under the framework of a "fixed bin" model or "adaptive bin" model. In general, mixtures of ith class-conditional multinomial c.d.f.'s are not identifiable because they are, in general, used to approximate ith class-conditional c.d.f.'s about which little is known a priori. We then ask what constraints must be imposed on the ith class-conditional c.d.f.'s approximating them, to insure identifiability? The following propositions give a partial answer to this question.

Let $X_s = \{x_{s_k}\}_1^v$ be a sequence of one dimensional samples where $x_{s_k} = 1$ with probability p^o and 0 with probability $1 - p^o$, with a single pattern class ω_i active for all v samples. Let $F(\{x_{s_k}\}_1^v | v, p^i, \omega_i)$ be the c.d.f. of the samples when class ω_i is active. The distribution of $\{x_{s_k}\}_1^v$ is thus a mixture c.d.f.; the corresponding parameter-conditional mixture c.d.f. is

$$F(\{X_{s_k}\}_1^v | B) = \sum_{i=1}^{M} F(\{X_{s_k}\}_1^v | v, p^i, \omega_i)\, P_i \tag{A-3}$$

The quenstion is when can p^i and P_i, $i = 1,2,\ldots,M$, be uniquely found given $F(\{x_{s_k}\}_1^v)$? The following Proposition A.4 by Teicher[10] gives sufficient conditions for a unique solution to exist for a more general problem than the one above. Proposition A.5 applies to the specific problem (A-3). Propositions A.6 and A.7 are extensions of Proposition A.5 to the multinomial case.

Proposition A.4. Let $\mathcal{F}_1 = \{F(x|n_i', p^{i'}),\ 1 \leqq i \leqq k'\}$ and $\mathcal{F}_2 = \{F(x|n'', p^{i''}),\ 1 \leqq i \leqq k''\}$ denote 2 finite families of binomial distributions; let k = number of elements in $\mathcal{F}_1 \cup \mathcal{F}_2$ and $\bar{n}_1 > \bar{n}_2 > \ldots > \bar{n}_h$ be the distinct integral parameters of the members of $\mathcal{F}_1 \cup \mathcal{F}_2$. A necessary but, in general, insufficient condition for

$$\text{(a)}\quad \sum_{i=1}^{k'} C'_i\, F(x|n_i', p^{i'}) \underset{x}{\equiv} \sum_{i=1}^{k''} C_i''\, F(x|n_i'', p^{i''}),\quad \sum_{i=1}^{k'} C_i' = \sum_{i=1}^{k''} C_i'' = 1,\ 0 < C_i', C_i''$$

to imply

(b) $k' = k''$, $(n_i', p^{i'}) = (n''_{j_i}, p^{j_i''})$ for some permutation $(j_1,\ldots,j_k)$ of $(1,2,\ldots,k)$ is that

(c) $\bar{n}_h \geq r_h - 1$

where r_i = number of occurrences of $\bar{n}_i$ among the elements of $\mathcal{F}_1 \cup \mathcal{F}_2$, $i \leq i \leq h$. A sufficient condition that (a) imply (b) is that (c) and (d) $\bar{n}_i - \bar{n}_{i+1} \geq r_i$, $1 \leq i \leq h-1$ hold

A special case of Proposition A.4 is $n_i = v$, $i = 1,2,\ldots,$. This corresponds to X_s always consisting of v samples, no matter what class is active. Teicher's[10] Proposition for this case is:

<u>Proposition A.5</u>. Let $\mathcal{F} = \{F(x|v, p_1^i), 0 < p_1^i < 1, i = 1,2,\ldots,M\}$ constitute a one-parameter family of binomial distributions, v being fixed. A necessary and sufficient condition that the class $U_{j=1}^M \mathcal{H}_j$ of all finite mixtures of at most M elements of $\mathcal{F}$ be identifiable is that $v \geq 2M - 1$.

The significance of Proposition A.5 is that p_1^1 and p_1^2 in Fig. 2 can be uniquely found if X_s consists of at least three samples from the same class. This may be a strong constraint, but there are some adaptive problems in practice where one class or pattern will be active long enough to take 2M - 1 samples, where M in the number of classes.

We will now give an extension of Proposition A.4 and Proposition A.5 to a parameter-conditional mixture of rth class-conditional <u>multinomial distributions</u>. Define a parameter-conditional mixture of multinomial c.d.f.'s, using (A-2), as

$$F(X_s|B) = \sum_{r=1}^{M} F(X_s|n_r, \{p_\xi^r\}_1^R) \qquad \text{(A-4)}$$

where the set $\{p_\xi^r\}_1^R$ are the R probabilities characterizing the rth class-conditional multinomial distribution. We state the following proposition and proof:

<u>Proposition A.6</u>. A sufficient condition for the multinomial family

$\mathcal{F} = \{F(X_s | n, \{p_\xi^r\}_1^R\}$ to give an identifiable class of mixtures is that

(a) $\bar{n}_h \geqq r_h - 1$ and

(b) $\bar{n}_i - \bar{n}_{i+1} \geqq r_i$, $1 \leqq i \leqq h - 1$ hold.

PROOF:

Let $p_1^i = p^i$

$$\sum_{\xi=2}^{R+1} p_\xi^i = 1 - p^i, \; i = 1,2,\ldots,M$$

Then (a) and (b) are sufficient, by Proposition A.4, for $\{p^i\}_1^M$ and $\{P_i\}_1^M$ to be uniquely found. In general, repeat the above with

$$p_\eta^i = p^i, \; \sum_{\xi \neq \eta} p_\xi^i = 1 - p^i, \qquad \begin{array}{l} i = 1,2,\ldots,M \\ \eta = 2,3,\ldots,R \end{array}$$

The following is a special case of Proposition A.6, as Proposition A.5 was a special case of Proposition A.4.

Proposition A.7. Let $\mathcal{F} = \{F(X_s | v, \{p_\xi^r\}_1^R, 0 < p_\xi^r < 1, r = 1,2,\ldots,M\}$ constitute a family of rth class-conditional multinomial distributions, v being fixed. A sufficient condition that the class $\bigcup_{j=1}^M \mathcal{H}_j$ of all finite mixtures of at most M elements of $\mathcal{F}$ be identifiable is that $v \geq 2M - 1$.

APPENDIX B

IMPLICIT SOLUTIONS FOR MAXIMUM LIKELIHOOD ESTIMATORS

For $f(x_s|\omega_i,B_i)$, B, and B_o given respectively by (5.6), (5.7), and (5.8), maximum likelihood estimators $\tilde{m}_1$, $\tilde{m}_2$, $\tilde{\sigma}$, and $\tilde{P}_1$ are obtained as follows:

$$F(\{x_s\}_1^n|B) = \prod_{s=1}^{n} f(x_s|B)$$

such that the likelihood function is

$$\log f(\{x_s\}_1^n|B) = \sum_{s=1}^{n} \log f(x_s|B)$$

Differentiating this likelihood function with respect to θ_i gives

$$T_i(\{x_s\}_1^n|B) = \sum_{s=1}^{n} \frac{\partial \log f(x_s|B)}{\partial\theta_i}$$

For later use, define

$$\psi_s = \tilde{P}_1 \exp\left[-\tfrac{1}{2} \frac{(x_s-\tilde{m}_1)^2}{\tilde{\sigma}^2}\right] + (1-P_1) \exp\left[-\tfrac{1}{2} \frac{(x_s-\tilde{m}_2)^2}{\tilde{\sigma}^2}\right]$$

$$\phi_{i,s} = \frac{\tilde{P}_i \exp\left[-\tfrac{1}{2} \frac{(x_s-\tilde{m}_i)^2}{\tilde{\sigma}^2}\right]}{\psi_s}$$

(a) For $\theta_i = m_i$, $i = 1,2$ we obtain

$$T_i(\{X_s\}_1^n|B) = \sum_{s=1}^{n} \frac{1}{f(x_s|B)} \frac{\partial f(x_s|B)}{\partial\theta_i} =$$

$$\sum_{s=1}^{n} \frac{(x_s-m_i)\,(P_i|\sigma^2)\,\exp\left[-\tfrac{1}{2} \frac{(x_s-m_i)^2}{\sigma^2}\right]}{P_i \exp\left[-\tfrac{1}{2}\frac{(x_s-m_o)^2}{\sigma^2}\right] + (1-P_i) \exp\left[-\tfrac{1}{2}\frac{(x_s-m_2)^2}{\sigma^2}\right]} \qquad \text{(B-1)}$$

$$= \frac{1}{\sigma^2} \sum_{s=1}^{n} (x_s-m_i)\,\phi_{i,s} = 0, \; i = 1,2$$

Thus,

$$\tilde{m}_i = \frac{\sum_{i=1}^{n} x_s \, \phi_{i,s}}{\sum_{i=1}^{n} \phi_{i,s}}, \quad i = 1,2$$

(b) For $\theta_3 = \sigma$ we obtain

$$T_3 \left(\{x_s\}_1^n | B\right) = \sum_{s=1}^{n} \frac{1}{f(x_s|B)} \frac{\partial f(x_s|B)}{\partial \theta_3} =$$

$$\sum_{s=1}^{n} \frac{-\frac{1}{\sigma} f(x_s|B) + \frac{1}{\sigma\sqrt{2\pi}}\left[\frac{P_1(x_s-m_1)^2}{\sigma^2} e^{-\frac{1}{2}\frac{(x_s-m_1)^2}{\sigma^2}} + \frac{P_2(x_s-m_2)^2}{\sigma^3} e^{-\frac{1}{2}\frac{(x_s-m_2)^2}{\sigma^2}}\right]}{f(x_s|B)}$$

$$= \frac{1}{\sigma} \sum_{s=1}^{n} - 1 + \frac{1}{\sigma^2} \frac{\left[P_1(x_s-m_1)^2 \exp(-\frac{1}{2}\frac{(x_s-m_1)^2}{\sigma^2}) + P_2(x_s-m_2)^2 \exp(-\frac{1}{2}\frac{(x_s-m_2)^2}{\sigma^2})\right]}{\left[P_1 \exp(-\frac{1}{2}\frac{(x_s-m_i)^2}{\sigma^2}) + P_2 \exp(-\frac{1}{2}\frac{(x_s-m_2)^2}{\sigma^2})\right]}$$

$$= \frac{1}{\sigma} \sum_{s=1}^{n} - 1 + \frac{1}{\sigma^2}\left[(x_s-m_i)^2 \, \phi_{1,s} + (x_s - m_2)^2 \, \phi_{2,s}\right] = 0$$

$$\therefore \quad \tilde{\sigma}^2 = \frac{1}{n} \sum_{s=1}^{n} \left[(x_s-\tilde{m}_1)^2 \, \phi_{1,s} + (x_s-\tilde{m}_2)^2 \, \phi_{2,s}\right] \qquad \text{(B.2)}$$

(c) For $\theta_4 = P_1$ we obtain

$$T_4 \left(\{x_s\}_1^n | B\right) = \sum_{s=1}^{n} \frac{1}{f(x_s|B)} \frac{\partial f(x_s|B)}{\partial \theta_4} =$$

$$\sum_{s=1}^{n} \frac{\exp\left[-\frac{1}{2}\frac{(x_s-m_1)^2}{\sigma^2}\right] - \exp\left[-\frac{1}{2}\frac{(x_s-m_2)^2}{\sigma^2}\right]}{P_1 \exp\left[-\frac{1}{2}\frac{(x_s-m_1)^2}{\sigma^2}\right] + (1-P_1) \exp\left[-\frac{1}{2}\frac{(x_s-m_2)^2}{\sigma^2}\right]} =$$

$$\sum_{s=1}^{n} \left[\frac{\phi_{1,s}}{P_1} - \frac{\phi_{2,s}}{(1-P_1)}\right] = 0$$

or

$$\frac{1}{P_1} \sum_{s=1}^{n} \phi_{1,s} - \frac{1}{(1-P_1)} \sum_{s=1}^{n} \phi_{2,s} = 0$$

or

$$\frac{(1-P_1) \sum_{s=1}^{n} \phi_{1,s} - P_1 \sum_{s=1}^{n} \phi_{2,s}}{P_1\ (1-P_1)} = 0$$

or

$$\sum_{s=1}^{n} \phi_{1,s} - P_1\left[\sum_{s=1}^{n} \phi_{1,s} + \sum_{s=1}^{n} \phi_{2,s}\right] = 0$$

such that

$$\tilde{P}_1 = \frac{\sum_{s=1}^{n} \phi_{1,s}}{\sum_{s=1}^{n} \phi_{1,s} + \sum_{s=1}^{n} \phi_{2,s}} = \sum_{s=1}^{n} \phi_{2,s} \tag{B-3}$$

The maximum likelihood estimators are summarized as follows:

$$\tilde{P}_1 = \sum_{s=1}^{n} \phi_{1,s},\ \tilde{P}_2 = 1-\tilde{P}_1 \tag{B-4}$$

$$\tilde{m}_i = \frac{\sum_{s=1}^{n} x_s\ \phi_{i,s}}{\sum_{s=1}^{n} \phi_{i,s}},\ i = 1,2 \tag{B-5}$$

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{s=1}^{n} \left[(x_s-\tilde{m}_1)^2\ \phi_{1,s} + (x_s-\tilde{m}_2)^2\ \phi_{2,s}\right] \tag{B-6}$$

where

$$\psi_s = \tilde{P}_1 \exp\left[-\tfrac{1}{2}\frac{(x_s-\tilde{m}_1)^2}{\tilde{\sigma}^2}\right] + (1-\tilde{P}_1) \exp\left[-\tfrac{1}{2} \frac{(x_s-\tilde{m}_2)^2}{\tilde{\sigma}^2}\right] \tag{B-7}$$

and

$$\phi_{i,s} = \frac{\tilde{P}_i \exp\left[-\tfrac{1}{2}\frac{(x_s-\tilde{m}_i)^2}{\tilde{\sigma}^2}\right]}{\psi_s} \tag{B-8}$$

APPENDIX C

REGULARITY CONDITIONS

First and Second Regularity Condition of $F(x|B)$ for $B = (m_1, m_2, \sigma, P_1)$

In this appendix we show that the first and second regularity conditions for $F(x|B)$ are satisfied for the binary, gaussian nonsupervisory problem.

The first regularity condition is verified by showing that (5.19) holds for $j = 1, 2, 3, 4$, corresponding to m_1, m_2, σ, and P_1 respectively; thus, four equations must be verified. The second regularity condition is verified by showing that (5.20) holds for all combinations of j and k, $j,k = 1, 2, 3, 4$; thus, sixteen equations must be verified. Because of symmetry, however, only three of the former and seven of the latter need be verified.

1) Let $\theta_4 = P_1$. Then

$$T_4(x|B) = \frac{\partial}{\partial P_1} \text{ Log } f(x|B) = \frac{\left[e^{-\frac{(x-m_1)^2}{2\sigma^2}} - e^{-\frac{(x-m_2)^2}{2\sigma^2}}\right]}{\left[P_1 e^{\frac{-(x-m_1)^2}{2\sigma^2}} + (1-P_1)\, e^{\frac{-(x-m_2)^2}{2\sigma^2}}\right]}$$

and

$$\int_{-\infty}^{\infty} T_4(x|B)\, f(x|B)\, dx = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty} \left[e^{\frac{-(x-m_1)^2}{2\sigma^2}} - e^{\frac{-(x-m_2)^2}{2\sigma^2}}\right] d\,x = 1 - 1 = 0$$

2) Let $\theta_i = m_i$, $i = 1, 2$. Then

$$T_i(x|B) = \frac{\partial}{\partial m_i} \log\,(x|B) = \frac{P_i \frac{(x-m_i)}{\sigma^2}\, e^{\frac{-(x-m_i)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma\, f(x|B)},\ i = 1,2$$

and

$$\int_{-\infty}^{\infty} T_i(x|B)\ f(x|B)\ d\,x = \frac{P_i}{\sigma^2} \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} (x-m_i)\ e^{-\frac{(x-m_i)^2}{2\sigma^2}}\ dx = 0$$

3) Let $\theta_3 = \sigma$

$$T_3\ (x|B) = \frac{\partial}{\partial\sigma} \log f(x|B) =$$

$$\frac{1}{\sigma}\left[-1+\frac{1}{\sigma^3\sqrt{2\pi}}\ \frac{P_1(x-m_1)^2\ e^{-\frac{1}{2}\frac{(x-m_1)^2}{\sigma^2}} + (1-P_1)\ (x-m_2)^2\ e^{-\frac{1}{2}\frac{(x-m_2)^2}{\sigma^2}}}{f(x|B)}\right]$$

and

$$\int_{-\infty}^{\infty} T_3\ (x|B)\ f(x|B)\ d\,x = -\frac{1}{\sigma} + \frac{1}{\sigma^3}\left[P_1\ \sigma^2 + (1-P_1)\ \sigma^2\right] = -\frac{1}{\sigma} + \frac{1}{\sigma} = 0$$

The seven equations for the second regularity conditions are shown to be satisfied as follows:

1) For $\theta_4 = P(\omega_1) = P_1$, we show $C_{44}(B) + D_{44}(B) = 0$:

$$T_4\ (x|B) = \frac{e^{-\frac{(x-m_1)^2}{2\sigma^2}} - e^{-\frac{(x-m_2)^2}{2\sigma^2}}}{P_1\left[e^{-\frac{(x-m_1)^2}{2\sigma^2}} - e^{-\frac{(x-m_2)^2}{2\sigma^2}}\right] + e^{-\frac{(x-m_2)^2}{2\sigma^2}}}$$

$$T_{44}(x|B) = -\frac{\left[e^{-\frac{(x-m_1)^2}{2\sigma^2}} - e^{\frac{(x-m_2)^2}{2\sigma^2}}\right]^2}{2\pi\ \sigma^2\ f^2(x|B)}$$

such that

$$D_{44}(B) = \frac{-1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} \frac{\left[e^{-\frac{(x-m_1)^2}{2\sigma^2}} - e^{-\frac{(x-m_2)^2}{2\sigma^2}}\right]}{f(x|B)\,\sqrt{2\pi}\,\sigma}\,dx$$

On the other hand,

$$C_{44}(B) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} \frac{\left[e^{-\frac{(x-m_1)^2}{2\sigma^2}} - e^{-\frac{(x-m_2)^2}{2\sigma^2}}\right]^2}{f(x|B)\,\sqrt{2\pi}\,\sigma}\,dx$$

Thus: $D_{44}(B) + C_{44}(B) = 0$

2) For $\theta_i = m_i$, $i = 1, 2$, we show $C_{ii}(B) + D_{ii}(B) = 0$

$$T_{ii}(x|B) =$$

$$\frac{P_i}{\sqrt{2\pi}\,\sigma}\left[\frac{-\frac{1}{\sigma^2}\,e^{-\frac{(x-m_i)^2}{2\sigma^2}} + \frac{(x-m_i)^2}{\sigma^4}\,e^{-\frac{(x-m_i)^2}{2\sigma^2}}}{f(x|B)} - \frac{\frac{(x-m_i)^2}{\sigma^4}\,e^{-\frac{(x-m_i)^2}{\sigma^2}}\,\frac{P_i}{\sqrt{2\pi}\,\sigma}}{f^2(x|B)}\right.$$

such that

$$D_{ii}(B) = \frac{-P_i}{\sigma^2\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty} e^{-\frac{(x-m_i)^2}{2\sigma^2}}\,dx + \frac{P_i}{\sigma^4\,2\pi\,\sigma}\int_{-\infty}^{\infty} (x-m_i)^2\,e^{-\frac{(x-m_i)^2}{2\sigma^2}}\,dx$$

$$- \frac{P_i^2}{\sqrt{2\pi}\,\sigma^5\,\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty}\frac{(x-m_i)^2}{f(x|B)}\,e^{-\frac{(x-m_i)^2}{\sigma^2}}\,dx =$$

$$-\frac{P_i}{\sigma^2} + \frac{P_i}{\sigma^2} - \frac{P_i^2}{2\pi\,\sigma^6}\int_{-\infty}^{\infty}\frac{(x-m_i)^2}{f(x|B)}\,e^{-\frac{(x-m_i)^2}{\sigma^2}}\,dx$$

On the other hand,

$$C_{ii}(B) = \frac{P_i^2}{2\pi\ \sigma^6} \int_{-\infty}^{\infty} \frac{(x-m_i)^2}{f(x|B)}\ e^{-\frac{(x-m_i)^2}{\sigma^2}}\ dx$$

Thus: $D_{ii}(B) + C_{ii}(B) = 0, \quad i = 1,2.$

3) For $\theta_3 = \sigma$:

$$T_{33}(x|B) = \frac{1}{\sigma^2} + \frac{\left[P_1 \frac{(x-m_1)^4}{\sigma^3}\ e^{-\frac{1}{2}\frac{(x-m_1)^2}{\sigma^2}} + (1-P_1)\frac{(x-m_2)^2}{\sigma^3}\ e^{-\frac{1}{2}\frac{(x-m_2)^2}{\sigma^2}}\right.}{\sigma^4\ \sqrt{2\pi}\ \ f(x|B)}$$

$$- \frac{4}{\sigma^5\sqrt{2\pi}}\ \frac{\left[P_1(x-m_1)^2\ e^{-\frac{1}{2}\frac{(x-m_1)^2}{\sigma^2}} + (1-P_1)\ (x-m_2)^2\ e^{-\frac{1}{2}\frac{(x-m_1)^2}{\sigma^2}}\right]}{f(x|B)}$$

$$- \frac{1}{\sigma^8\ 2\pi}\ \frac{\left[P_1(x-m_1)^2\ e^{-\frac{1}{2}\frac{(x-m_1)^2}{\sigma^2}} + (1-P_1)\ (x-m_2)^2\ e^{-\frac{1}{2}\frac{(x-m_2)^2}{\sigma^2}}\right]^2}{f^2(x|B)}$$

such that

$$D_{33}(B) = \frac{1}{\sigma^2} + \frac{3}{\sigma^2} - \frac{4}{\sigma^2}$$

$$- \frac{1}{\sigma^8\ 2\pi} \int_{-\infty}^{\infty} \frac{\left[P_1\ (x-m_1)^2\ e^{-\frac{1}{2}\frac{(x-m_1)^2}{\sigma^2}} + (1-P_1)\ (x-m_2)^2\ e^{-\frac{1}{2}\frac{(x-m_2)^2}{\sigma^2}}\right]^2}{f(x|B)}\ dx$$

On the other hand,

$$C_{33}(B) = \frac{1}{\sigma^2} - \frac{1}{\sigma^2} + \frac{1}{\sigma^8\ 2\pi} \int_{-\infty}^{\infty} \frac{\left[P_1(x-m_1)^2\ e^{-\frac{1}{2}\frac{(x-m_1)^2}{\sigma^2}} + (1-P_1)(x-m_2)^2\ e^{-\frac{1}{2}\frac{(x-m_2)^2}{\sigma^2}}\right]^2}{f(x|B)} dx$$

Thus: $D_{33}(B) + C_{33}(B) = 0$.

We now consider the four remaining "cross" regularity conditions.

4) To show $D_{ij}(B) + C_{ij}(B) = 0$, $i,j = 1, 2$, $i \neq j$:

Define

$$E_i = \frac{P_i}{\sqrt{2\pi}} \frac{(x-m_i)}{\sigma^3} e^{-\frac{(x-m_i)^2}{2\sigma^2}}$$

Then

$$D_{ij} = -\int_{-\infty}^{\infty} \frac{E_i\ E_j}{f(x|B)}\, dx$$

On the other hand,

$$C_{ij}(B) = \int_{-\infty}^{\infty} \frac{E_i E_j}{f(x|B)}\, dx$$

Thus: $D_{ij}(B) + C_{ij}(B) = 0$, $i,j = 1, 2$, $i \neq j$

5) To show $D_{i4} + C_{i4} = 0$, $i = 1, 2$:

$$T_{i4}(x|B) = \frac{(x-m_i)\ e^{-\frac{(x-m_i)^2}{2\sigma^2}}}{\sqrt{2\pi}\ \sigma^3\ f(x|B)} = \frac{T_i(x|B)}{f(x|B)} \frac{\partial f(x|B)}{\partial P_i}$$

Define $J_i = e^{-\frac{(x-m_i)^2}{2\sigma^2}}$

Then

$$T_{i4}(x|B) = \frac{(x-m_i)\ J_i}{\sqrt{2\pi}\ \sigma^3\ f(x|B)} - \frac{P_i(x-m_i)\ J_i}{f^2(x|B)\ \sqrt{2\pi}\ \sigma^3} \quad \frac{(J_i-J_j)}{\sqrt{2\pi}\ \sigma} =$$

$$\frac{(x-m_i)\ J_i}{\sqrt{2\pi}\ \sigma^3\ f(x|B)} \left[1 - \frac{(J_i-J_j)\ P_i}{f(x|B)\ \sqrt{2\pi}\ \sigma}\right]$$

such that

$$D_{i4} = 0 - \frac{P_i}{2\pi\ \sigma^4} \int_{-\infty}^{\infty} \frac{(x-m_i)\ J_i(J_i-J_j)}{f(x|B)}\ dx,\ i = 1,2$$

On the other hand,

$$T_i(x|B)\ T_4(x|B) = \left[\frac{P_i(x-m_i)J_i}{\sqrt{2\pi}\ \sigma^3\ f(x|B)}\right]\left[\frac{(J_i-J_j)}{f(x|B)\ \sqrt{2\pi}\ \sigma}\right],\ i = 1,2$$

and

$$C_{i4}(B) = \frac{P_i}{2\pi\ \sigma^4} \int_{-\infty}^{\infty} \frac{(x-m_i)\ J_i\ (J_i-J_j)}{f(x|B)}\ dx$$

Thus: $D_{i4}(B) + C_{i4}(B) = 0,\ i = 1,\ 2$

Finally, one can show

6) $D_{34}(B) + C_{34}(B) = 0$

and

7) $D_{i3}(B) + C_{i3}(B) = 0,\ i = 1,\ 2.$