

AD-A055 997

RICE UNIV HOUSTON TEX DEPT OF ELECTRICAL ENGINEERING  
COMPUTATIONALLY EFFICIENT ESTIMATORS FOR THE BAYES RISK.(U)

F/G 5/8

UNCLASSIFIED

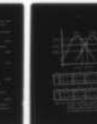
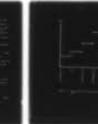
EE-TR-7804

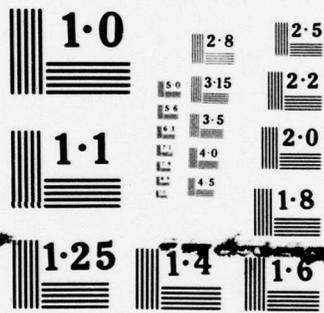
AFOSR-TR-78-1081

AFOSR-75-2777

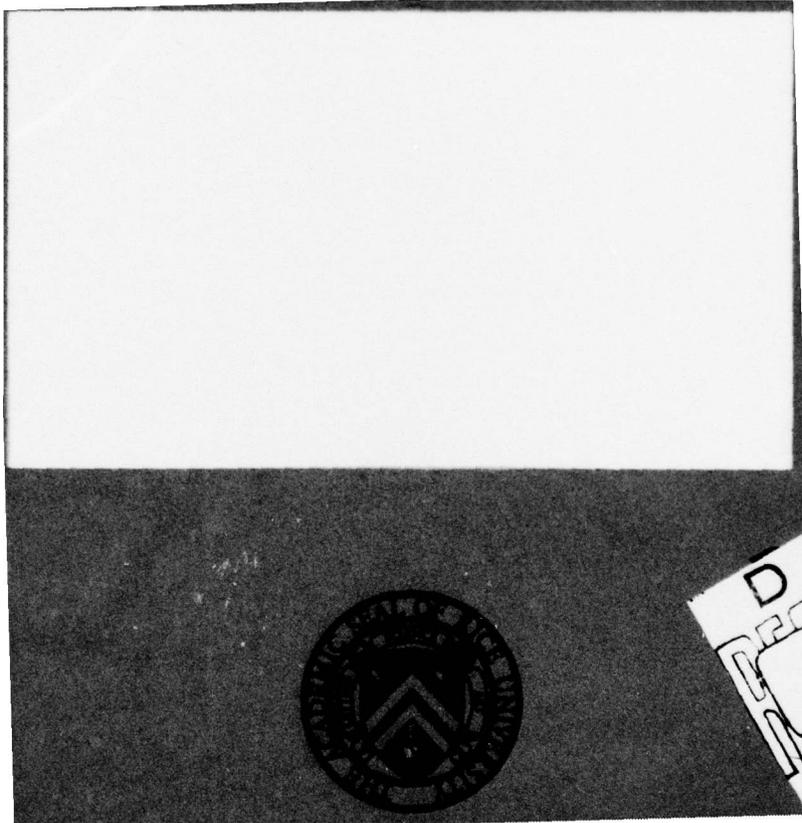
NL

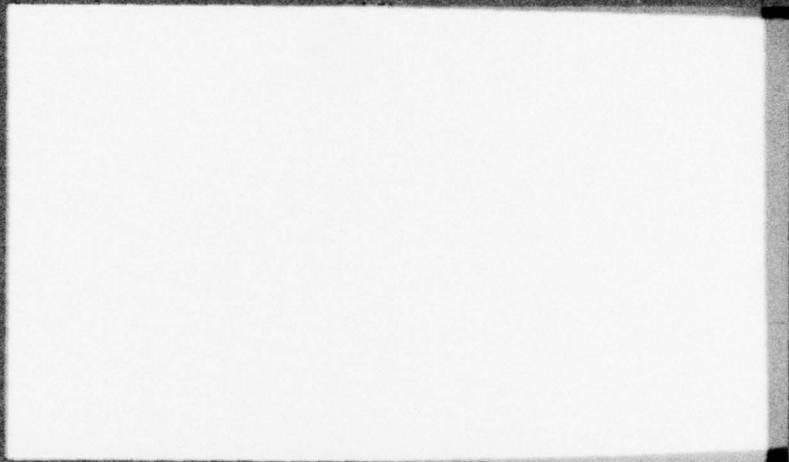
1 of 2  
ADA  
055997





NATIONAL BUREAU OF STANDARDS  
MICROCOPY RESOLUTION TEST CHART





ARMY FORCE OFFICE OF SCIENTIFIC RESEARCH (AFOSR)  
NOTICE OF TRANSMITTAL TO DDC  
This technical report has been reviewed and is  
approved for public release LAW AFR 100-12 (7b).  
Distribution is unlimited.  
A. B. BLOOM  
Technical Information Officer

AD A 055997

C

6 COMPUTATIONALLY EFFICIENT ESTIMATORS FOR THE BAYES RISK.

by

10 Lynn D. Wilcox\* and Rui J.P. de Figueiredo†

\*Department of Mathematical Sciences

†Department of Electrical Engineering

Rice University, Houston, Texas 77001

May, 1978

9 TECHNICAL REPORT, # 7804 ✓

12 112P/

11 May 78

DDC  
JUL 5 1978  
RESERVE

14 EE-TR-7804

15 AFOSR-75-2777

AD No. /  
DDC FILE COPY

This document has been approved for public release and sale; its distribution is unlimited.

18 AFOSR

19 TR-78-1081

This work was supported by the AFOSR Grant 75-2777. ✓

78 06 27 081  
403 244

JOB

COMPUTATIONALLY EFFICIENT ESTIMATORS  
FOR THE BAYES RISK

Lynn D. Wilcox

ABSTRACT

A computationally efficient estimator for the Bayes risk is one which achieves a desired accuracy with a minimum of computation. In many problems, for example speech recognition, point evaluations of the class conditional densities are computationally costly. Density evaluations are the single most important factor contributing to the computational effort in Bayes risk estimation, thus the amount of computation required by a Bayes risk estimator is defined as the average number of conditional density evaluations it performs. The accuracy of a risk estimator is defined by its variance. *→ next page*

Existing estimators for the Bayes risk, namely the error count estimator and the posterior estimator, require for each sample  $X_j$ ,  $j=1,2,\dots,N$ , evaluation of the class conditional density  $f_m(X_j)$  for each class  $m=1,2,\dots,M$ , a total of  $N \cdot M$  density evaluations. For problems such as speech recognition, where the number of classes  $M$  is large and density evaluations costly, these estimators are impractical from a computational aspect.

A new class of estimators of the general form  $\hat{R}(T)$  is proposed. An estimator  $\hat{R}(T)$  is defined by associating with each class  $m$  a subset  $T_m$  of the  $M$  classes. For two classes, only the error count and posterior estimators belong to this class. For more than two classes, several new estimators for the Bayes risk are included.

Estimators requiring fewer density evaluations are derived from the class of estimators of the general form  $\hat{R}(T)$  as follows. A scalar para-

meter  $\alpha$  determines the sets  $T_m(\alpha)$  of classes that are " $\alpha$ -close" to each class  $m$ , hence an estimator  $\hat{R}(\alpha)$  of the general form  $\hat{R}(T)$ . As  $\alpha$  varies, the sets  $T_1(\alpha), \dots, T_M(\alpha)$  vary and a family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  of risk estimators is achieved. Each estimator in the family is characterized by the average number of density evaluations it requires and its variance.

The optimal estimator from the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  is defined as that estimator with maximum computational efficiency, where the computational efficiency of an estimator is the inverse of the product of the average number of density evaluations it requires and its variance. The optimal estimator requires the least amount of computation to achieve a given accuracy, or, symmetrically, achieves the greatest accuracy with a minimum of computation.

In practice, the true optimal estimator cannot be determined since this would in effect require knowledge of the true risk  $R$ . Thus a technique whereby the first  $n$  of the total  $N$  samples are used to approximate the optimal estimator is proposed. The  $n$  samples should contain enough information on the closeness of the classes to determine an almost optimal estimator. The last  $N-n$  samples are used in the approximate optimal estimator to obtain an accurate estimate of the risk with a minimum of computation.

ACCESSION FOR	
NTIS	White Section <input checked="" type="checkbox"/>
DOC	Buff Section <input type="checkbox"/>
UNANNOUNCED	<input checked="" type="checkbox"/>
JUSTIFICATION	
DISTRIBUTION/AVAILABILITY CODES	
SPECIAL	
A	

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 78-1081 ✓	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle)  COMPUTATIONALLY EFFICIENT ESTIMATORS FOR THE BAYES RISK	5. TYPE OF REPORT & PERIOD COVERED  Interim	
	6. PERFORMING ORG. REPORT NUMBER EE7804	
7. AUTHOR(s)  Lynn D. Wilcox and Rui J.P. de Figueiredo	8. CONTRACT OR GRANT NUMBER(s)  AFOSR 75-2777	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Rice University Department of Electrical Engineering Houston, Texas 77001	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS  61102F 2304/A2	
11. CONTROLLING OFFICE NAME AND ADDRESS  Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332	12. REPORT DATE May 1978	
	13. NUMBER OF PAGES 87	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report)  UNCLASSIFIED	
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
16. DISTRIBUTION STATEMENT (of this Report)  Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of this abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)  Pattern recognition; Bayes risk; error estimation		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)  A computationally efficient estimator for the Bayes risk is one which achieves a desired accuracy with a minimum of computation. Existing estimators based on error count or the risk function require, at each sample of the test data set, point evaluation of the class conditional density for each of the classes. In problems such as speech recognition, where the number of classes is large and point evaluations of the densities complex, these estimators are impractical from a computational aspect.		

DD FORM 1 JAN 73 1473

UNCLASSIFIED

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract

A new family of estimators for the Bayes risk is defined. Computational forms for estimators in the family reduce the number of densities that must be evaluated at each test sample. Thus a computationally efficient estimator may be chosen from the family.

UNCLASSIFIED

## TABLE OF CONTENTS

	Page
1. INTRODUCTION TO THE RESEARCH TOPIC	1
1.1 Introduction	1
1.2 Review of Previous Work	3
1.3 Approach and Development in the Present Work	6
2. BASIC CONCEPTS ASSOCIATED WITH THE BAYES RISK	10
3. PROPOSED NEW ESTIMATORS FOR THE BAYES RISK	14
3.1 Introduction	14
3.2 Estimators Based on Unrestricted Sampling	14
3.2.1 Remarks on Error Count and Posterior Estimators	15
3.2.2 A General Form for Bayes Risk Estimators	17
3.2.3 A Parameterized Family of Estimators for the Bayes Risk	26
3.2.4 Computational Requirements for Estimators in the Family	31
3.2.5 Variances of Estimators in the Family	40
3.2.6 Examples	43
3.3 Estimators Based on Stratified Sampling	45
3.3.1 A Parameterized Family of Bayes Risk Estimators	46
3.3.2 Variances of Estimators in the Family	48
3.3.3 Computational Requirements for Estimators in the Family	51

	Page
4. OPTIMAL ESTIMATORS	53
4.1 Introduction	53
4.2 Computational Efficiency: A Criterion for the Optimal Estimator	55
4.3 An Algorithm for Maximization of the Computational Efficiency	57
4.4 Comparison of the Optimal Estimator with the Error Count and Posterior Estimators	63
4.5 Approximation of the Optimal Estimator	68
4.6 Examples	75
5. CONCLUSIONS	80
5.1 Summary of Results	80
5.2 Recommendations for Further Research	83
BIBLIOGRAPHY	85
APPENDIX	A-1
A. Data From Example 1	A-1
B. Data From Example 2	B-1

## CHAPTER 1

## INTRODUCTION TO THE RESEARCH TOPIC

1.1 Introduction

The task of a pattern recognition system is to decide to which of  $M$  classes a given pattern belongs. The decision is made on the basis of a set of measurements  $X$  taken on the pattern and is specified by the decision rule  $\delta(X)$ . The performance of the system may be characterized by the probability that it makes a classification error. The decision rule which minimizes the probability of classification error is called the Bayes rule and the resulting minimum probability of classification error the Bayes risk.

The Bayes risk represents the optimal performance of a pattern recognition system for a given set of measurements  $X$ . As such it may be regarded as the intrinsic difficulty of the problem, or the confusability of the  $M$  classes. Suppose one wanted to compare the difficulty of two speech recognition tasks. The number of words in each vocabulary would be one criterion. However, one should also consider the confusability of the words in each vocabulary, as measured by the Bayes risk for each task.

In this thesis, we study estimators for the Bayes risk in terms of the amount of computation they require and their accuracy. It is assumed that the class conditional densities  $f_1(x), \dots, f_M(x)$  and priors  $\pi_1, \dots, \pi_M$  are known so that attention may be focused on the actual forms for risk estimators. The results will also apply asymptotically if the unknown densities are estimated on training data which is independent of the

test data used in the risk estimators, provided the density estimates are asymptotically unbiased and consistent.

In many problems, point evaluations of the class conditional densities are computationally costly. For example, in speech recognition [23, 1], the class conditional density  $f_m(x)$  would be the probability that the output phone string  $x$  was caused by the  $m^{\text{th}}$  word in the vocabulary. Evaluation of  $f_m(x)$  involves determining all phonetic realizations of the  $m^{\text{th}}$  word, and for each phonetic realization, all segmentation and classification errors that would result in the output phone string  $x$ . In estimation of the Bayes risk, density evaluations are the single most important factor contributing to the computational effort. Thus the amount of computation required by a Bayes risk estimator is defined as the average number of class conditional density evaluations involved in the estimation procedure.

Existing estimators for the Bayes risk, namely the error count estimator [6] and the posterior estimator [11], require for each sample  $X_j$ ,  $j=1,2,\dots,N$  in the test data set, evaluation of the class conditional densities  $f_1(X_j), \dots, f_M(X_j)$ , a total of  $N \times M$  density evaluations. Thus for problems such as speech recognition, where the number of classes  $M$  is large and density evaluations costly, these estimators are impractical from a computational aspect.

We propose several new estimators for the Bayes risk. In particular, a family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  of unbiased and consistent risk estimators, indexed on the scalar parameter  $\alpha$ , is defined. The parameter  $\alpha$  determines, for each sample  $X_j$ , the classes  $\mathcal{L}$  for which the class conditional density  $f_{\mathcal{L}}(X_j)$  must be evaluated in forming the estimator  $\hat{R}(\alpha)$ . In

general,  $\hat{R}(\alpha)$  may be computed with fewer density evaluations than the  $NXM$  required by the existing estimators. Bayes risk estimators are evaluated in terms of their computational efficiency, defined as the inverse of the product of their variance times the average number of density evaluations they require. An estimator with maximum computational efficiency is considered optimal. The optimal estimator has the property that a minimum of computation is required to achieve a given accuracy.

### 1.2 Review of Previous Work

The usual test data for estimation of the Bayes risk is a sample of measurements or observations  $X$  and their true classifications or labels  $\theta$ . This type of sample will be referred to as unrestricted [31, 19], since the statistician has no control over the label of a sample. There are two existing forms for Bayes risk estimators: the error count estimator and the posterior estimator. The error count estimator [19, 6] is simply the proportion of samples  $X$  whose classification by the Bayes rule disagrees with its true classification  $\theta$ . The posterior estimator was first suggested by Chow [3], later formalized by Fukunaga and Kessel [11] and discovered independently by Lissack and Fu [27]. It is the sample mean of the risk function evaluated at the sample points. It is interesting that the posterior estimator ignores information on the class labels, yet has a lower variance than the error count estimator [11].

Another sampling technique called stratified sampling is often possible [31]. As opposed to unrestricted sampling, the statistician chooses a priori a class label and samples observations  $X$  with that label. By choosing the number of samples per class appropriately, the variance

of a given estimator may be reduced. Neyman [31] determines the optimal number of samples per class by minimizing the variance of the estimator. Highleyman [19] applied the stratified sampling technique to the error count estimator. He did not choose the optimal sample sizes, but rather chose the number of samples per class as proportional to the prior probability of that class. He shows that even this heuristic choice achieves a reduction in the variance of the error count estimator.

Moore, Whitsitt and Landgrebe [30] later applied stratified sampling to the posterior estimator. They show the heuristic sample size is not optimal, but the optimal sample sizes are impractical since they depend on unknown variances. Stratified sampling with sample sizes proportional to the priors also reduce the variance of the posterior estimator. Moore, Whitsitt and Landgrebe [30] give the interesting result that while for unrestricted sampling, the posterior estimator has smaller variance than the error count, this is not necessarily true when a stratified sample is used, even with the optimal choice of sample sizes.

Both the error count and posterior estimators for the Bayes risk require knowledge of the class conditional densities  $f_m(x)$ ,  $m=1,2,\dots,M$ . When these densities are unknown, one way to proceed is to estimate the densities and use the estimates in the estimators as if they were the true densities. Cover and Wagner [4] call these two-step procedures.

When the test data used for the risk estimator must also be used to estimate the densities (i.e. when the test data is the same as the training data), the question of data use must be considered. If the samples used in the density estimates are also used in the risk estimator, an optimistic bias in the resulting estimate for the Bayes risk is observed.

If the data set is large, an alternative is to partition the data and use part to estimate the densities and the rest in the estimator. Highleyman [19] tried to optimize this partition but Kanal and Chandrasekaran [25] questioned his assumptions. The leave-one-out technique of Lachenbruch and Mickey [26] attempts to remove bias by estimating the densities on all but one sample and using the deleted sample in the estimator for the Bayes risk. Each sample in turn is left out and the resulting risk estimate is the average of the one-point estimates. An excellent discussion of these and other methods of data use is given in Toussaint [37] and Kanal [24].

Several density estimates have been considered for use in Bayes risk estimators. Lissack and Fu [27] and Fukunaga and Kessel [12] assume a parametric form for the densities (exponential family and Gaussian respectively) and estimate the parameters. Fukunaga and Kessel [12], Fralick and Scott [9], and Whitsitt and Landgrebe [39] use Parzen estimators [32]. Fukunaga and Kessel [15] and Fukunaga and Hostetter [13] consider nearest neighbor techniques for direct estimation of the risk function used in the posterior estimator. Lissack and Fu [27] apply Loftsgaarden and Quesenberry [29] nearest neighbor density estimates to obtain estimates for the class posterior probabilities. A good discussion of results when various combinations of estimator form, data use and density estimates are tried is given in Whitsitt and Landgrebe [39].

Computational difficulties in Bayes risk estimators arise from the fact that for each sample  $X$ , the conditional density  $f_m(X)$  of the sample  $X$  given class  $m$  must be evaluated for all classes  $m=1,2,\dots,M$ . Whitsitt and

Landgrebe [39] consider this problem when the densities are estimated with Parzen estimators using a Gaussian kernel. They propose an edited Parzen estimator for the densities  $f_m(x)$ . Rather than averaging the kernel over all data points with class label  $m$ , the average is taken over only those data points labeled  $m$  which are the  $k$  nearest neighbors to the point  $X$ .

Any density estimate which requires nearest neighbors may be simplified by algorithms which find nearest neighbors efficiently. These include condensed nearest neighbor rules by Hart [18] and Swonger [36], a branch and bound algorithm by Fukunaga and Narendra [14] and preprocessing techniques by Fisher and Patrick [8], Yunk [40], and Friedman et al. [10].

The above techniques achieve reduction in computation by simplifying the evaluation of the conditional densities  $f_m(X)$  at the data points. In this thesis, computationally efficient estimators are achieved by reducing the number of densities which must be evaluated at a given sample point. Thus rather than evaluate  $f_m(X)$  for all classes  $m=1,2,\dots,M$  at the point  $X$ , we might only evaluate  $f_m(X)$  for  $m$  in a subset of the total classes. This is profitable in problems such as speech recognition where the number of classes  $M$  is large and computation of conditional densities complex [23, 1].

### 1.3 Approach and Development in the Present Work

A new class of Bayes risk estimators of the general form  $\hat{R}(T)$  is proposed. The estimator  $\hat{R}(T)$  is defined on the basis of sets  $T_1, \dots, T_M$ , where  $T_m$  is a set of classes associated with class  $m$ . Subject to mild restrictions, any choice of the sets  $T_1, \dots, T_M$  results in an unbiased, consistent Bayes risk estimator. Both of the existing estimators, namely the error count and the posterior, belong to the class of estimators of

the general form  $\hat{R}(T)$ .

In order to obtain risk estimators which require fewer class conditional density evaluations, we restrict the set  $T_m$  of classes associated with class  $m$  as follows. A scalar parameter  $\alpha$  determines the set of classes  $T_m(\alpha)$  that are " $\alpha$ -close" to class  $m$ , that is, a sample  $X$  whose true classification  $\theta$  is  $m$  is likely, as determined by  $\alpha$ , to be classified as  $i$ , whenever classes  $i$  and  $m$  are " $\alpha$ -close". As  $\alpha$  varies, the set of classes  $T_1(\alpha), \dots, T_M(\alpha)$  vary and a family of risk estimators  $\{R(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$ , indexed on the parameter  $\alpha$ , is achieved.

The definition of the sets  $T_1(\alpha), \dots, T_M(\alpha)$  allows the estimator  $\hat{R}(\alpha)$  to be formed with fewer class conditional density evaluations. Thus rather than evaluate the conditional density  $f_{\ell}(X_j)$  at each sample  $X_j$ ,  $j=1,2,\dots,N$  for each class  $\ell=1,2,\dots,M$ , the estimator  $\hat{R}(\alpha)$  requires evaluation of  $f_{\ell}(X_j)$  for only those classes  $\ell$  in a subset  $Q_{\theta_j}(\alpha)$  of the total classes  $\{1,2,\dots,M\}$ , whenever the joint density  $f_{\theta_j}(X_j)\pi_{\theta_j}$  of the sample  $X_j$  and its class label  $\theta_j$  is greater than  $\alpha$ . The subset  $Q_{\theta_j}(\alpha)$  is the set of classes that are " $\alpha$ -close" to each class that is " $\alpha$ -close" to the class label  $\theta_j$  of  $X_j$ .

The amount of computation required by the estimator  $\hat{R}(\alpha)$  is expressed by  $NXC(\alpha)$ , the average number of conditional densities that must be evaluated, where  $N$  is the sample size and  $C(\alpha)$  is the average number of conditional densities per sample used in forming  $\hat{R}(\alpha)$ . The error count and posterior estimators require all  $M$  conditional densities per sample, a total of  $M \times N$  density evaluations. Thus if  $\alpha$  is such that  $C(\alpha)$  is much smaller than  $M$ , the estimator  $\hat{R}(\alpha)$  would be computationally preferable to

either of the existing estimators.

The accuracy of an estimator  $\hat{R}(\alpha)$  based on  $N$  samples is given by its variance  $V(\alpha)/N$ . Thus the larger the sample size  $N$ , or the smaller the coefficient of variance  $V(\alpha)$ , the more accurate the estimator. The estimator in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  with the smallest coefficient of variance  $V(\alpha)$  has the property that it requires the least number of samples  $N$  to achieve a given accuracy. However, the size of the sample is not sufficient to characterize the amount of computation required by an estimator in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$ , since the average number of density evaluations per sample  $C(\alpha)$  required by each estimator must be considered.

We define the computational efficiency  $\mathcal{CE}(\alpha)$  of an estimator  $\hat{R}(\alpha)$  as the inverse of the product of its variance and the average number of density evaluations it requires, thus  $\mathcal{CE} = 1/V(\alpha) \times C(\alpha)$ . The optimal estimator  $\hat{R}(\alpha^*)$  from the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  is determined by choosing  $\alpha^*$  to maximize the computational efficiency  $\mathcal{CE}(\alpha)$ . The optimal estimator  $\hat{R}(\alpha^*)$  has the property that it achieves a given accuracy with a minimum of computation [16, 17], or symmetrically, that for a given amount of computation,  $\hat{R}(\alpha^*)$  is the most accurate estimator for the Bayes risk  $R$ .

In practice, the optimal estimator could not be determined in this way since this would in effect require knowledge of the true risk  $R$ . Thus a technique is proposed whereby a subset  $n$  of the total  $N$  samples is used to approximate the optimal estimator. The number  $n$  of samples should contain enough information on the closeness of the classes to determine an almost optimal estimator. The remaining  $N-n$  samples are

used in the approximate optimal estimator to obtain an accurate estimate of the risk with a minimum of computation.

## CHAPTER 2

## BASIC CONCEPTS ASSOCIATED WITH THE BAYES RISK

A general pattern recognition system may be modeled mathematically in terms of a probability triple  $(\Omega, F, P)$ , an observation random variable  $X$ , and a labeling random variable  $\theta$ . Let  $\Omega$  be the space of patterns  $\omega$ ,  $F$  a sigma field of subsets of  $\Omega$  and  $P$  a probability measure defined on  $F$ . The patterns  $\omega \in \Omega$  are to be classified into one of  $M$  classes, where the classes  $H_1, H_2, \dots, H_M$  are a disjoint partition of  $\Omega$ . If a pattern  $\omega \in H_m$  we say  $\omega$  is in class  $m$ .

The random variable  $\theta : \Omega \rightarrow \{1, 2, \dots, M\}$  specifies the class of a pattern  $\omega$ , so that  $\theta(\omega) = m$  whenever  $\omega \in H_m$ .  $\theta$  is referred to as the class label or simply the label of a pattern. The prior probability of the  $m^{\text{th}}$  class is given by

$$\pi_m = P[H_m] = P[\theta=m]. \quad (2-1)$$

In practice, the patterns  $\omega \in \Omega$  are not actually observed. Rather, one observes a set of measurements made on  $\omega$ . The random variable  $X : \Omega \rightarrow S \subset R^d$  specifies the measurements  $X(\omega) = x \in S$  made on a pattern  $\omega$ . Assume the conditional density of  $X$  given  $\theta=m$  exists and is continuous and denote it by  $f_m(x)$ . Then the unconditional density of  $X$ , or the mixture density is given by

$$f(x) = \sum_{l=1}^M \pi_l f_l(x). \quad (2-2)$$

Also, the posterior probability of class  $m$ , the probability that  $\theta=m$  given the observation  $X(\omega)=x$  is

$$p_m(x) = \frac{f_m(x)\pi_m}{f(x)} \quad (2-3)$$

On the basis of the observation  $X(\omega)=x$ , the recognition system tries to decide the true classification of the pattern  $\omega$ , i.e. the value of  $\theta(\omega)$ . This decision may be specified by a behavioral decision rule  $\delta(x) = (\delta_1(x), \delta_2(x) \dots \delta_M(x))$ , where  $\delta_m(x)$  is the probability that the recognition system classifies a pattern  $\omega$  as belonging to class  $m$ , given the observation  $x$ . Thus  $\delta_m(x) \geq 0$ ,  $m=1,2,\dots,M$  and

$$\sum_{m=1}^M \delta_m(x) = 1.$$

Given a decision rule  $\delta$ , the probability  $R(\delta)$  that the system makes a classification error may be written

$$\begin{aligned} R(\delta) &= \sum_{m=1}^M \pi_m \int_S \sum_{i \neq m} \delta_i(x) f_m(x) dx \\ &= \sum_{m=1}^M \pi_m \int_S (1 - \delta_m(x)) f_m(x) dx \end{aligned} \quad (2-4)$$

It is well known [2, 7] that the decision rule  $\delta^*$  which minimizes the probability of classification error  $R(\delta)$  is the Bayes decision rule  $\delta^*$ , where ties are broken at random and

$$\delta_m^*(x) = \begin{cases} 1 & f_m(x)\pi_m > f_\ell(x)\pi_\ell \quad \forall \ell=1,2,\dots,M, \ell \neq m \\ 0 & \exists k \neq m \ni f_k(x)\pi_k > f_m(x)\pi_m \end{cases} \quad (2-5)$$

The minimum probability of classification error resulting from Bayes decision rule is called Bayes risk and is denoted by  $R$ .

The error function  $\mathcal{E}_\theta(X)$  is defined for  $\theta=m$ ,  $X=x$  as one minus the Bayes rule  $\delta_m^*(x)$ ,

$$\mathcal{E}_m(x) = \begin{cases} 0 & f_m(x)\pi_m > f_l(x)\pi_l \quad \forall l=1,2,\dots,M, l \neq m \\ 1 & \exists k \neq m \ni f_k(x)\pi_k > f_m(x)\pi_m \end{cases} \quad (2-6)$$

Then the bayes risk R is

$$R = \sum_{m=1}^M \pi_m \int_S \mathcal{E}_m(x) f_m(x) dx \quad (2-7)$$

Note that R is just the expectation, over the random variables X and  $\theta$ , of the error function  $\mathcal{E}_\theta(X)$ , so

$$R = E\{\mathcal{E}_\theta(X)\} \quad (2-8)$$

The conditional risk  $R_m$  is the probability of classification error given class m,

$$R_m = \int_S \mathcal{E}_m(x) f_m(x) dx \quad (2-9)$$

Thus  $R_m$  is the conditional expectation of the error function  $\mathcal{E}_\theta(X)$  given  $\theta=m$ .

$$R_m = E\{\mathcal{E}_\theta(X) | \theta=m\} \quad (2-10)$$

Since

$$R = E\{\mathcal{E}_\theta(X)\} = E\{E\{\mathcal{E}_\theta(X) | \theta\}\} \quad (2-11)$$

we have that

$$R = \sum_{m=1}^M \pi_m E\{\mathcal{E}_\theta(X) | \theta=m\} = \sum_{m=1}^M \pi_m R_m \quad (2-12)$$

The risk function  $r(x)$  is the probability of classification error given the observation  $X=x$ . Symbolically,

$$r(x) = \sum_{m=1}^M \mathcal{E}_m(x) p_m(x) = \sum_{m=1}^M \mathcal{E}_m(x) \frac{f_m(x) \pi_m}{f(x)} \quad (2-13)$$

Thus  $r(x)$  is the conditional expectation of the error function given  $X=x$ ,

$$r(x) = E\{\mathcal{E}_\theta(X) | X=x\} \quad (2-14)$$

Then the Bayes risk is the expectation over  $X$  of the risk function  $r(X)$ , since

$$R = E\{\mathcal{E}_\theta(X)\} = E\{E\{\mathcal{E}_\theta(X) | X\}\} = E\{r(X)\} \quad (2-15)$$

## CHAPTER 3

## PROPOSED NEW ESTIMATORS FOR THE BAYES RISK

3.1 Introduction

In this section, a general form  $\hat{R}(T)$  for Bayes risk estimators is defined. Based on the general form, a family of estimators  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$ , indexed on a scalar parameter  $\alpha$ , is derived. A computational form for estimators in the family is given which in general allows these estimators to be computed on the basis of fewer density evaluations per sample. The computational requirements of an estimator  $\hat{R}(\alpha)$  may be described by the expected number of density evaluations  $C(\alpha)$  per sample. The behavior  $C(\alpha)$  as a function of  $\alpha$ , as well as the behavior of the variance  $V(\alpha)$  of  $\hat{R}(\alpha)$  are discussed.

Two sampling techniques for estimation of the Bayes risk are considered: unrestricted sampling and stratified sampling. The basic difference between these sampling techniques is that in unrestricted sampling, the number of samples with a given class label is random, while for stratified sampling the statistician chooses a priori the number of samples with a given class label.

3.2 Estimators Based on Unrestricted Sampling

For unrestricted sampling, the data is a set sequence  $\{(X_1, \theta_1), (X_2, \theta_2), \dots, (X_N, \theta_N)\}$  of  $N$  independent random vectors identically distributed as  $(X, \theta)$ . The joint density of  $(X, \theta)$  at  $X=x, \theta=m$  is given by  $f_m(x)\pi_m$ , where  $f_m(x)$  is the conditional density of  $X$  given the class label  $\theta=m$  and  $\pi_m$  is the prior probability of class  $m$ . The marginal density of the observation  $X$  at  $X=x$  is  $f(x)$ , the mixture density.

The proportion of samples  $X_j$  whose class label  $\theta_j$  is  $m$  is random, with mean  $\pi_m$ .

### 3.2.1 Remarks on Error Count and Posterior Estimators

The error count estimator  $\hat{R}(ec)$  for the Bayes risk  $R$  is formed by counting the proportion of samples  $X_j$  whose classification by the Bayes rule disagrees with the true class label  $\theta_j$ . Symbolically,

$$\hat{R}(ec) = \frac{1}{N} \sum_{j=1}^N \mathcal{E}_{\theta_j}(X_j) \quad (3-1)$$

The error count estimator is unbiased, since

$$E\{\hat{R}(ec)\} = E\{\mathcal{E}_{\theta}(X)\} = R. \quad (3-2)$$

It is also consistent [19, 6], since

$$\text{VAR}\{\hat{R}(ec)\} = \frac{1}{N} \text{VAR}\{\mathcal{E}_{\theta}(X)\} = \frac{R(1-R)}{N} \quad (3-3)$$

The error count estimator  $\hat{R}_m(ec)$  for the conditional risk  $R_m$  given class  $m$  is

$$\hat{R}_m(ec) = \frac{1}{N} \sum_{j=1}^N I_m(\theta_j) \frac{\mathcal{E}_m(X_j)}{\pi_m} \quad (3-4)$$

$$\text{where } I_m(\theta) = \begin{cases} 1, & \theta=m \\ 0, & \theta \neq m \end{cases}$$

$\hat{R}_m(ec)$  is an unbiased estimator for  $R_m$  since [33]

$$E\{\hat{R}_m(ec)\} = \frac{E\{I_m(\theta)\mathcal{E}_m(X)\}}{\pi_m} = E\{\mathcal{E}_{\theta}(X) | \theta=m\} = R_m \quad (3-5)$$

Note that the error count estimator  $\hat{R}_m(ec)$  considers only classification errors made on samples  $X_j$  whose class labels  $\theta_j = m$ . Also,

$$\hat{R}(ec) = \sum_{m=1}^M \pi_m \hat{R}_m(ec) . \quad (3-6)$$

The posterior estimator  $\hat{R}(p)$  for the Bayes risk  $R$  is the sample mean of the risk function  $r(X_j)$  over the samples  $X_j$   $j=1,2 \dots N$  [11].

Thus

$$\hat{R}(p) = \frac{1}{N} \sum_{j=1}^N r(X_j) = \frac{1}{N} \sum_{j=1}^N \sum_{m=1}^M \mathcal{E}_m(X_j) \frac{f_m(X_j) \pi_m}{f(X_j)} \quad (3-7)$$

The posterior estimator is unbiased, since

$$E\{\hat{R}(p)\} = E\{r(X)\} = R. \quad (3-8)$$

It is also consistent, since

$$\text{VAR}\{\hat{R}(p)\} = \frac{1}{N} \text{VAR}\{r(X)\} = \frac{1}{N} \left[ \int_S r^2(x) f(x) dx - R^2 \right] \quad (3-9)$$

It has been shown [11] that the posterior estimator has smaller variance than the error count estimator. This follows from the fact that since  $0 \leq r(x) \leq 1 - \frac{1}{M}$

$$\int_S r^2(x) f(x) dx \leq R - \frac{R}{M} \quad (3-10)$$

and thus

$$\text{VAR}\{\hat{R}(p)\} \leq \frac{R(1-R)}{N} - \frac{R}{MN} \leq \frac{R(1-R)}{N} = \text{VAR}\{\hat{R}(ec)\} \quad (3-11)$$

The posterior estimator  $\hat{R}_m(p)$  for the conditional risk  $R_m$  is defined by

$$\hat{R}_m(p) = \frac{1}{N} \sum_{j=1}^N \mathcal{E}_m(X_j) \frac{f_m(X_j)}{f(X_j)} \quad (3-12)$$

In contrast to the error count estimator, the posterior estimator  $\hat{R}_m(p)$  considers errors made on all samples  $X_j$ , regardless of their class labels  $\theta_j$ . In fact, the posterior estimator makes no use of the class labels. The expectation of  $\hat{R}_m(p)$  is thus taken with respect to the mixture density  $F(x)$ , so

$$\begin{aligned} E\{\hat{R}_m(p)\} &= E\left\{\mathcal{E}_m(X) \frac{f_m(X)}{f(X)}\right\} \\ &= \int_S \mathcal{E}_m(x) \frac{f_m(x)}{f(x)} f(x) dx = \int_S \mathcal{E}_m(x) f_m(x) dx = R_m \end{aligned} \quad (3-13)$$

Thus  $\hat{R}_m(p)$  is an unbiased estimator of the conditional risk  $R_m$ . Again

$$\hat{R}(p) = \sum_{m=1}^M \pi_m \hat{R}_m(p) \quad (3-14)$$

### 3.2.2 A General Form for Bayes Risk Estimators

The error count estimator for the conditional risk  $R_m$  in effect considers only those samples  $X_j$  whose class labels  $\theta_j$  are equal to  $m$ . The posterior estimator for  $R_m$  considers all samples, regardless of their true classification. This concept may be generalized by associating with each class  $m$  a subset  $T_m$  of the total classes, and forming an estimator  $\hat{R}_m(T)$  based on those samples  $X_j$  whose class labels  $\theta_j$  are elements of  $T_m$ .

Specifically, for each  $m=1,2,\dots,M$ , let  $T_m = \{i_1, i_2, \dots, i_{p_m}\}$  be a set of  $p_m$  classes associated with class  $m$ , where  $i_j$ ,  $j=1,2,\dots,p_m$  are members of the set. The sets  $T_m$ ,  $m=1,2,\dots,M$  may be chosen arbitrarily,

TABLE 3-1

$T_1 = \{1,2,3\}$	$Q_1 = \{1,2,3\}$	$T_1 = \{1,3\}$	$Q_1 = \{1,3\}$
$T_2 = \{1,2,3\}$	$Q_2 = \{1,2,3\}$	$T_2 = \{2\}$	$Q_2 = \{2\}$
$T_3 = \{1,2,3\}$	$Q_3 = \{1,2,3\}$	$T_3 = \{1,3\}$	$Q_3 = \{1,3\}$
$T_1 = \{1,3\}$	$Q_1 = \{1,2,3\}$	$T_1 = \{1\}$	$Q_1 = \{1\}$
$T_2 = \{2,3\}$	$Q_2 = \{1,2,3\}$	$T_2 = \{2,3\}$	$Q_2 = \{2,3\}$
$T_3 = \{1,2,3\}$	$Q_3 = \{1,2,3\}$	$T_3 = \{2,3\}$	$Q_3 = \{2,3\}$
$T_1 = \{1,2,3\}$	$Q_1 = \{1,2,3\}$	$T_1 = \{1,2\}$	$Q_1 = \{1,2\}$
$T_2 = \{1,2\}$	$Q_2 = \{1,2,3\}$	$T_2 = \{1,2\}$	$Q_2 = \{1,2\}$
$T_3 = \{1,3\}$	$Q_3 = \{1,2,3\}$	$T_3 = \{3\}$	$Q_3 = \{3\}$
$T_1 = \{1,2\}$	$Q_1 = \{1,2,3\}$	$T_1 = \{1\}$	$Q_1 = \{1\}$
$T_2 = \{1,2,3\}$	$Q_2 = \{1,2,3\}$	$T_2 = \{2\}$	$Q_2 = \{2\}$
$T_3 = \{2,3\}$	$Q_3 = \{1,2,3\}$	$T_3 = \{3\}$	$Q_3 = \{3\}$

All possible choices of the sets  $T_m$ ,  $m=1,2,3$  subject to restrictions (r1) and (r2) and the resulting sets  $Q_m$ ,  $m=1,2,3$ .

subject to the following restrictions.

$$(r1) \quad m \in T_m \quad \forall m=1,2,\dots,M$$

$$(r2) \quad i \in T_m \quad \text{iff} \quad m \in T_i \quad \forall i,m=1,2,\dots,M$$

Restriction (r1) requires that each class be associated with itself, and restriction (r2) requires that a class  $i$  be associated with class  $m$  if and only if class  $m$  is associated with class  $i$ .

For  $M$  classes, there are  $2^{\frac{M(M-1)}{2}}$  different ways to choose the sets  $T_m$ ,  $m=1,2,\dots,M$ , subject to restrictions (r1) and (r2). Table 3-1 lists the 8 choices for the case of  $M = 3$  classes.

The general form  $\hat{R}_m(T)$  for an estimator of the conditional risk  $R_m$  is defined by considering those samples  $X_j$  whose class labels  $\theta_j$  are in  $T_m$ . Thus

$$\hat{R}_m(T) = \frac{1}{N} \sum_{j=1}^N I_{T_m}(\theta_j) \mathcal{E}_m(X_j) \frac{f_m(X_j)}{\sum_{\ell \in T_m} f_\ell(X_j) \pi_\ell} \quad (3-15)$$

$$\text{where} \quad I_{T_m}(\theta_j) = \begin{cases} 1 & \theta_j \in T_m \\ 0 & \theta_j \notin T_m \end{cases}$$

Subject to the restriction (r1),  $\hat{R}_m(T)$  is an unbiased estimator for  $R_m$  for any choice of the set  $T_m$ , since

$$\begin{aligned}
E\{\hat{R}_m(T)\} &= E\left\{I_{T_m}(\theta) \frac{e_m(x)f_m(x)}{\sum_{\ell \in T_m} f_\ell(x)\pi_\ell}\right\} \\
&= \sum_{i=1}^M \int_S I_{T_m}(i) \frac{e_m(x)f_m(x)f_i(x)\pi_i}{\sum_{\ell \in T_m} f_\ell(x)\pi_\ell} dx \\
&= \int_S e_m(x)f_m(x) \frac{\sum_{i \in T_m} f_i(x)\pi_i}{\sum_{\ell \in T_m} f_\ell(x)\pi_\ell} dx \quad (3-16) \\
&= \int_S e_m(x)f_m(x) dx = R_m
\end{aligned}$$

A general estimator  $\hat{R}(T)$  for the unconditional risk is formed

as

$$\begin{aligned}
\hat{R}(T) &= \sum_{i=1}^M \pi_m \hat{R}_m(T) \\
&= \frac{1}{N} \sum_{j=1}^N \sum_{m=1}^M I_{T_m}(\theta_j) \frac{e_m(X_j)f_m(X_j)\pi_m}{\sum_{\ell \in T_m} f_\ell(X_j)\pi_\ell} \quad (3-17)
\end{aligned}$$

By linearity of the expectation operator,  $\hat{R}(T)$  is unbiased for any choice of the sets  $T_m$ ,  $m=1,2,\dots,M$  (subject to (r1)). By restriction (r2),  $I_{T_m}(\theta) = I_{T_\theta}(m)$ , thus from (3-17),  $\hat{R}(T)$  may be written

$$\hat{R}(T) = \frac{1}{N} \sum_{j=1}^N \sum_{m \in T_\theta_j} e_m(X_j) \frac{f_m(X_j)\pi_m}{\sum_{\ell \in T_m} f_\ell(X_j)\pi_\ell} \quad (3-18)$$

The general form  $\hat{R}(T)$  is also consistent for any restricted choice of the sets  $T_m$ ,  $m=1,2,\dots,M$ . This follows because

$$\begin{aligned} \text{VAR}\{\hat{R}(T)\} &= \frac{1}{N} \text{VAR}\left\{ \sum_{m \in T_\theta} \mathcal{E}_m(X) \frac{f_m(X)\pi_m}{\sum_{\ell \in T_m} f_\ell(X)\pi_\ell} \right\} \\ &= \frac{1}{N} \left\{ \sum_{i=1}^M \pi_i \int_S \left( \sum_{m \in T_i} \frac{\mathcal{E}_m(x) f_m(x)\pi_m}{\sum_{\ell \in T_m} f_\ell(x)\pi_\ell} \right)^2 f_i(x) dx - R^2 \right\} \end{aligned} \quad (3-19)$$

Note that when each class  $m$  is associated only with itself, that is when  $T_m = \{m\} \forall m=1,2,\dots,M$ , the general estimator  $\hat{R}(T)$  is just the error count estimator  $\hat{R}(ec)$ , since in this case

$$\begin{aligned} \hat{R}(T) &= \frac{1}{N} \sum_{j=1}^N \sum_{m=1}^M I_{T_m}(\theta_j) \mathcal{E}_m(X_j) \\ &= \frac{1}{N} \sum_{j=1}^N \mathcal{E}_{\theta_j}(X_j) = \hat{R}(ec) \end{aligned} \quad (3-20)$$

When each class  $m$  is associated with all other classes, that is when  $T_m = \{1,2,\dots,M\} \forall m=1,2,\dots,M$  then the posterior estimator  $\hat{R}(p)$  is obtained, since

$$\begin{aligned} \hat{R}(T) &= \frac{1}{N} \sum_{j=1}^N \sum_{m=1}^M I_{T_m}(\theta_j) \frac{\mathcal{E}_m(X_j) f_m(X_j)\pi_m}{\sum_{\ell=1}^M f_\ell(X_j)\pi_\ell} \\ &= \frac{1}{N} \sum_{j=1}^N \sum_{m=1}^M \frac{\mathcal{E}_m(X_j) f_m(X_j)\pi_m}{f(X_j)} = \hat{R}(p) \end{aligned} \quad (3-21)$$

The number of different estimators for the Bayes risk specified by the general form  $\hat{R}(T)$  in (3-18) is  $2^{\frac{M(M-1)}{2}}$ , the number of different ways to choose the sets  $T_m$ ,  $m=1,2,\dots,M$ , subject to the restrictions (r1) and (r2). If the number of classes  $M = 2$ , only two estimators may be obtained, namely the error count and posterior. For  $M > 2$ ,

the general form  $\hat{R}(T)$  specifies several new estimators, depending on how the sets  $T_m$ ,  $m=1,2,\dots,M$  are chosen. Let us first consider how to choose the sets  $T_m$ ,  $m=1,2,\dots,M$  so that an estimator  $\hat{R}(T)$  with minimum variance is achieved.

It was shown in section 3.2.1 that the posterior estimator  $\hat{R}(p)$  has smaller variance than the error count estimator  $\hat{R}(ec)$ . The following theorem generalizes this result by showing that the choice of the sets  $T_m$ ,  $m=1,2,\dots,M$  which minimizes the variance of the estimator  $\hat{R}(T)$  is  $T_m^* = \{1,2,\dots,M\}$   $\forall m=1,2,\dots,M$ . But  $\hat{R}(T^*)$  is just the posterior estimator  $\hat{R}(p)$ , thus the posterior estimator has the smallest variance of any estimator of the general form  $\hat{R}(T)$ .

Theorem 3-1

Let  $\hat{R}(T)$  be the general estimator for the Bayes risk given by equation (3-18), with the sets  $T_m$ ,  $m=1,2,\dots,M$  chosen arbitrarily. Let  $\hat{R}(T^*)$  be the general estimator with the sets chosen as

$$T_m^* = \{1,2,\dots,M\} \quad \forall m=1,2,\dots,M$$

$$\text{Then } \text{VAR}\{\hat{R}(T^*)\} \leq \text{VAR}\{\hat{R}(T)\}$$

Proof:

$$\text{Let } r_T(X, \theta) = \sum_{m \in T_\theta} e_m(X) \frac{f_m(X) \pi_m}{\sum_{l \in T_m} f_l(X) \pi_l} \quad (3-22)$$

Then from equation (3-19)

$$\text{VAR}\{\hat{R}(T)\} = \frac{1}{N} \text{VAR}\{r_T(X, \theta)\} \quad (3-23)$$

The conditional expectation of  $r_T(X, \theta)$ , given  $X$ , is  $r(X)$ , the risk function. To see this,

$$\begin{aligned} E\{r_T(X, \theta) | X\} &= \sum_{i=1}^M r_T(X, i) p_i(X) \\ &= \sum_{i=1}^M \sum_{m \in T_i} \mathcal{E}_m(X) \frac{f_m(X) \pi_m}{\sum_{\ell \in T_m} f_\ell(X) \pi_\ell} p_i(X) \end{aligned} \quad (3-24)$$

Now  $p_i(X) = \frac{f_i(X) \pi_i}{f(X)}$  from equation (2-3),

and by restriction (r2),  $\sum_{i=1}^M \sum_{m \in T_i} = \sum_{m=1}^M \sum_{i \in T_m}$ . Thus from (3-24),

$$\begin{aligned} E\{r_T(X, \theta) | X\} &= \sum_{m=1}^M \sum_{i \in T_m} \frac{\mathcal{E}_m(X) f_m(X) \pi_m}{\sum_{\ell \in T_m} f_\ell(X) \pi_\ell} \frac{f_i(X) \pi_i}{f(X)} \\ &= \sum_{m=1}^M \mathcal{E}_m(X) \frac{f_m(X) \pi_m}{f(X)} = r(X) \end{aligned} \quad (3-25)$$

Also since  $\hat{R}(T^*) = \hat{R}(p)$ ,

$$\text{VAR}\{\hat{R}(T^*)\} = \frac{1}{N} \text{VAR}\{r(X)\} \quad (3-26)$$

Now the variance of the conditional expectation is less than the total variance, since by [34]

$$\begin{aligned} &\text{VAR}\{E\{r_T(X, \theta) | X\}\} \\ &= \text{VAR}\{r_T(X, \theta)\} - E\{\text{VAR}\{r_T(X, \theta) | X\}\} \\ &\leq \text{VAR}\{r_T(X, \theta)\} \end{aligned} \quad (3-27)$$

$$\begin{aligned} \text{Thus } \text{VAR}\{\hat{R}(T^*)\} &= \frac{1}{N} \text{VAR}\{r(X)\} \\ &= \frac{1}{N} \text{VAR}\{E\{r_T(X, \theta) | X\}\} \leq \frac{1}{N} \text{VAR}\{r_T(X, \theta)\} = \text{VAR}\{\hat{R}(T)\} \end{aligned}$$

If the sets  $T_m$ ,  $m=1, 2, \dots, M$  are chosen to minimize the variance of the estimator  $\hat{R}(T)$ , the resulting estimator is  $\hat{R}(T^*) = \hat{R}(p)$ , the posterior estimator. Another consideration in the choice of the sets  $T_m$ ,  $m=1, 2, \dots, M$  is the amount of computation required by the estimator  $\hat{R}(T)$ .

From equation (3-21),  $\hat{R}(p)$  requires at each sample  $X_j$ ,  $j=1, 2, \dots, N$  the computation of the conditional density  $f_\ell(X_j)$  for each class  $\ell=1, 2, \dots, M$ , a total of  $M \times N$  conditional density evaluations. In problems such as speech recognition [23, 1], where the number of classes  $M$  is large and evaluation of the conditional densities complex, the amount of computation required by the posterior estimator  $\hat{R}(p)$  is considerable. Thus we consider choosing the sets  $T_m$ ,  $m=1, 2, \dots, M$  in such a way that  $\hat{R}(T)$  may be computed on the basis of fewer density evaluations per sample.

For now let us disregard the fact that the error function  $\mathcal{E}_m(X_j)$  must be determined at each point  $X_j$  and for each class  $m \in T_{\theta_j}$ . Then from (3-18) the densities explicitly required in the estimator  $\hat{R}(T)$  at the point  $X_j$  are  $f_\ell(X_j)$  for all classes  $\ell$  in  $T_m$ , for all  $m$  in  $T_{\theta_j}$ . Define the sets  $Q_m$ ,  $m=1, 2, \dots, M$  associated with given sets  $T_m$ ,  $m=1, 2, \dots, M$  as the union over  $q$  in  $T_m$  of  $T_q$ .

Definition 3-1  $Q_m = \bigcup_{q \in T_m} T_q, m=1,2,\dots,M$

Table 3-1 gives the sets  $Q_m, m=1,2,\dots,M$  resulting from each choice of the sets  $T_m, m=1,2,\dots,M$ . Thus  $\hat{R}(T)$  for a given choice of sets  $T_m$  requires explicitly the evaluation of  $f_\ell(X_j) \forall \ell \in Q_{\theta_j}$  for each sample  $X_j, j=1,2,\dots,N^*$ .

Example 3-1 Suppose  $\hat{R}(T)$  is based on one sample  $(X_1, \theta_1)$  and that  $\theta_1 = 1$ . Then if  $T_1 = \{1,2\}$   $T_2 = \{1,2\}$  and  $T_3 = \{3\}$ .

$$\hat{R}(T) = \mathcal{E}_1(X_1) \frac{f_1(X_1)\pi_1}{f_1(X_1)\pi_1 + f_2(X_1)\pi_2} + \mathcal{E}_2(X_1) \frac{f_2(X_1)\pi_2}{f_1(X_1)\pi_1 + f_2(X_1)\pi_2}$$

Since  $Q_i = T_i \forall i$  and  $3 \notin Q_1$ ,  $f_3(X_1)$  is not used explicitly in  $\hat{R}(T)$ .

Of course, the error function  $\mathcal{E}_\theta(X)$  is an implicit function of all the conditional densities  $f_\ell(X), \ell = 1,2,\dots,M$ , and from (3-18)

$\mathcal{E}_m(X_j)$  must be computed  $\forall m \in T_{\theta_j}$ . In the next section, the set of classes  $T_m$  associated with class  $m$  will be chosen in such a way that

$\mathcal{E}_m(X_j), m \in T_{\theta_j}$ , may be determined on the basis of the densities

$f_\ell(X_j), \ell \in Q_{\theta_j}$  explicitly required by the estimator  $\hat{R}(T)$ . Thus estimators requiring fewer density evaluations are achieved.

---

\*

This analysis assumes we must always compute  $f_{\theta_j}(X_j)$ , even though this

is unnecessary when  $Q_{\theta_j} = \{\theta_j\}$ , since we would be evaluating 1 by  $\frac{f_{\theta_j}(X_j)}{f_{\theta_j}(X_j)}$ .

However,  $f_{\theta_j}(X_j)$  will always be needed to compute  $\mathcal{E}_m(X_j)$ .

### 3.2.3 A Parameterized Family of Estimators For the Bayes Risk

In section 3.2.2, it was shown that any choice of the sets  $T_m$  associated with class  $m$  would determine an unbiased, consistent estimator for the Bayes risk of the general form  $\hat{R}(T)$ , provided the sets  $T_m$ ,  $m=1,2,\dots,M$ , satisfy restrictions (r1) and (r2). In this section, we restrict the choice of the sets as follows. A scalar parameter  $\alpha \geq 0$  defines the set  $T_m(\alpha)$  of classes " $\alpha$ -close" to class  $m$ . Basically, class  $i$  is " $\alpha$ -close" to class  $m$  if the Bayes rule is likely (as determined by  $\alpha$ ) to classify a sample  $X$  whose true class label  $\theta=m$ , as class  $i$ . As  $\alpha$  varies, the sets  $T_m(\alpha)$  vary and a family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  of unbiased estimators is achieved. The definition of " $\alpha$ -closeness" allows the estimator  $\hat{R}(\alpha)$  to be computed with fewer density evaluations per sample.

Let  $\alpha \geq 0$  be a scalar and define  $T_m(\alpha)$ , the set of classes  $\alpha$ -close to class  $m$  by

#### Definition 3-2

$$T_m(\alpha) = \{i \mid \exists x \ni f_i(x)\pi_i > \alpha \text{ and } f_m(x)\pi_m > \alpha\}, \text{ for } m=1,2,\dots,M.$$

It follows from the definition that  $i \in T_m(\alpha)$  if and only if  $m \in T_i(\alpha)$ , thus restriction (r2) is met automatically for all  $\alpha$ . Restriction (r1), that  $m \in T_m(\alpha) \forall m=1,2,\dots,M$  is met by restricting  $0 \leq \alpha < \alpha_{\max}$ , where we define  $\alpha_{\max}$  as follows.

#### Definition 3-3

$$\alpha_{\max} = \min_{1 \leq l \leq M} \max_{x \in S} f_l(x)\pi_l.$$

A given  $\alpha$ ,  $0 \leq \alpha < \alpha_{\max}$  does not uniquely determine the sets  $T_m(\alpha)$ ,  $m=1,2,\dots,M$  since it is possible that  $\alpha \neq \alpha'$  but  $T_m(\alpha) = T_m(\alpha') \forall m=1,2,\dots,M$ . Let  $\alpha_{t_0}, \alpha_{t_1}, \dots, \alpha_{t_K}$  be the set of  $\alpha$ 's that in-

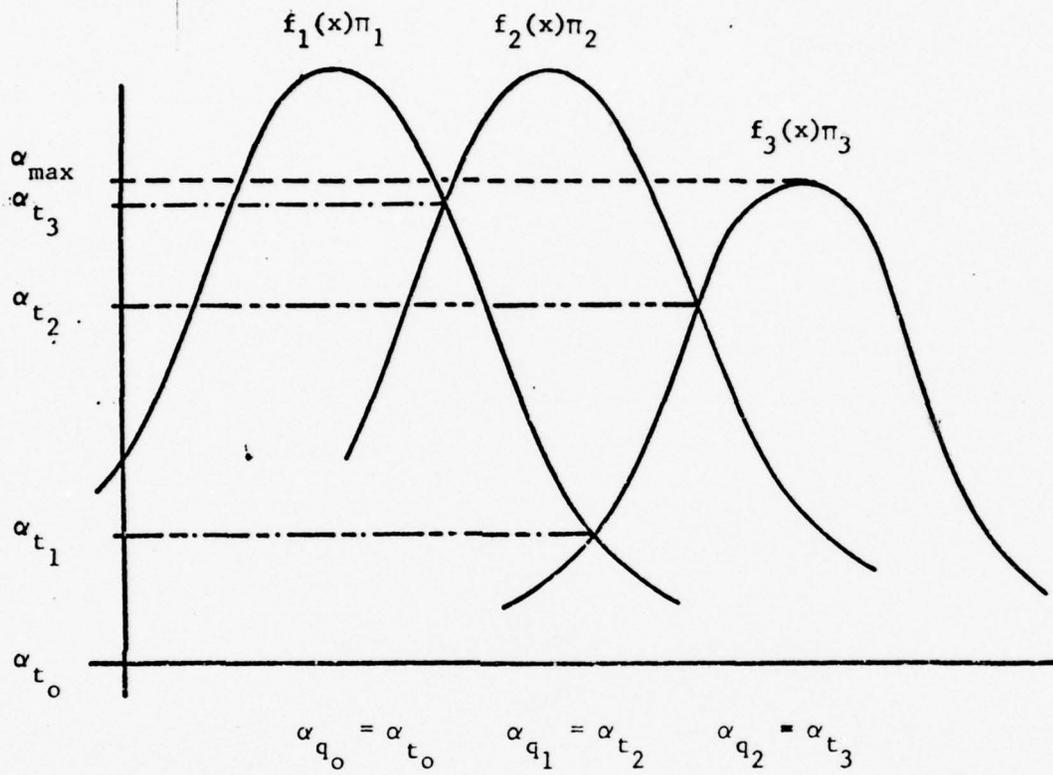


Figure 3-1. The points  $\alpha_{t_0} \dots \alpha_{t_3}$  which induce changes in the sets  $T_m$  and points  $\alpha_{q_0} \dots \alpha_{q_2}$  which induce changes in the sets  $Q_m$ .

TABLE 3-2

$0 \leq \alpha < \alpha_{t_1}$	$T_1(\alpha) = \{1, 2, 3\}$ $T_2(\alpha) = \{1, 2, 3\}$ $T_3(\alpha) = \{1, 2, 3\}$	$Q_1(\alpha) = \{1, 2, 3\}$ $Q_2(\alpha) = \{1, 2, 3\}$ $Q_3(\alpha) = \{1, 2, 3\}$
$\alpha_{t_1} \leq \alpha < \alpha_{t_2}$	$T_1(\alpha) = \{1, 2\}$ $T_2(\alpha) = \{1, 2, 3\}$ $T_3(\alpha) = \{2, 3\}$	$Q_1(\alpha) = \{1, 2, 3\}$ $Q_2(\alpha) = \{1, 2, 3\}$ $Q_3(\alpha) = \{1, 2, 3\}$
$\alpha_{t_2} \leq \alpha < \alpha_{t_3}$	$T_1(\alpha) = \{1, 2\}$ $T_2(\alpha) = \{1, 2\}$ $T_3(\alpha) = \{3\}$	$Q_1(\alpha) = \{1, 2\}$ $Q_2(\alpha) = \{1, 2\}$ $Q_3(\alpha) = \{3\}$
$\alpha_{t_3} \leq \alpha < \alpha_{\max}$	$T_1(\alpha) = \{1\}$ $T_2(\alpha) = \{2\}$ $T_3(\alpha) = \{3\}$	$Q_1(\alpha) = \{1\}$ $Q_2(\alpha) = \{2\}$ $Q_3(\alpha) = \{3\}$

Sets  $T_m(\alpha)$  and  $Q_m(\alpha)$  for all  $0 \leq \alpha < \alpha_{\max}$ .

duce changes in the sets  $T_m$  for some  $m$ , defined recursively as follows.

Definition 3-4

Let  $\alpha_{t_0} = 0$

Do  $i=0$  by 1 while  $\alpha_{t_i} < \alpha_{\max}$

Let  $\alpha_{t_{i+1}}$  be the smallest value of  $\alpha > \alpha_{t_i}$  such that

$T_m(\alpha_{t_{i+1}}) \neq T_m(\alpha_{t_i})$  for some  $m=1,2,\dots,M$ .

End.

Let  $\alpha_{t_K}$  be the largest value so defined.

Figure 3-1 shows three joint densities and the values  $\alpha_{t_0}, \alpha_{t_1}, \alpha_{t_2}, \alpha_{t_3}$  and  $\alpha_{\max}$ . Table 3-2 shows all possible sets  $T_m(\alpha), m=1,2,3$  that are defined. The  $K+1$  values  $\alpha_{t_0}, \alpha_{t_1}, \dots, \alpha_{t_K}$  determine the  $K+1$  possible sets  $T_m(\alpha)$ . Note that  $K \leq \frac{M(M-1)}{2}$ , so that the number of possible sets  $T_m$  de-

defined by  $\alpha$  is much less than the number of  $2^{\frac{M(M-1)}{2}}$  that were possible in section 3.2.2.

The parameter  $\alpha$  determines for each  $m=1,2,\dots,M$ , the sets  $T_m(\alpha)$  of classes  $\alpha$ -close to class  $m$ . An unbiased estimator  $\hat{R}_m(\alpha)$  for the conditional risk  $R_m$  based on samples  $X_j$  whose labels  $\theta_j$  are elements of  $T_m(\alpha)$  is defined from the general estimator  $\hat{R}_m(T)$  in (3-15) as

$$\hat{R}_m(\alpha) = \frac{1}{N} \sum_{j=1}^N I_{T_m(\alpha)}(\theta_j) \mathcal{E}_m(X_j) \frac{f_m(X_j)}{\sum_{\ell \in T_m(\alpha)} f_\ell(X_j) \pi_\ell} \quad (3-28)$$

The estimator  $\hat{R}(\alpha)$  for the unconditional risk  $R$  determined by the sets  $T_m(\alpha) m=1,2,\dots,M$  resulting from  $\alpha$  is, as in (3-18)

$$\hat{R}(\alpha) = \frac{1}{N} \sum_{j=1}^N \sum_{m \in T_{\theta_j}(\alpha)} \mathcal{E}_m(X_j) \frac{f_m(X_j) \pi_m}{\sum_{\ell \in T_m(\alpha)} f_\ell(X_j) \pi_\ell} \quad (3-29)$$

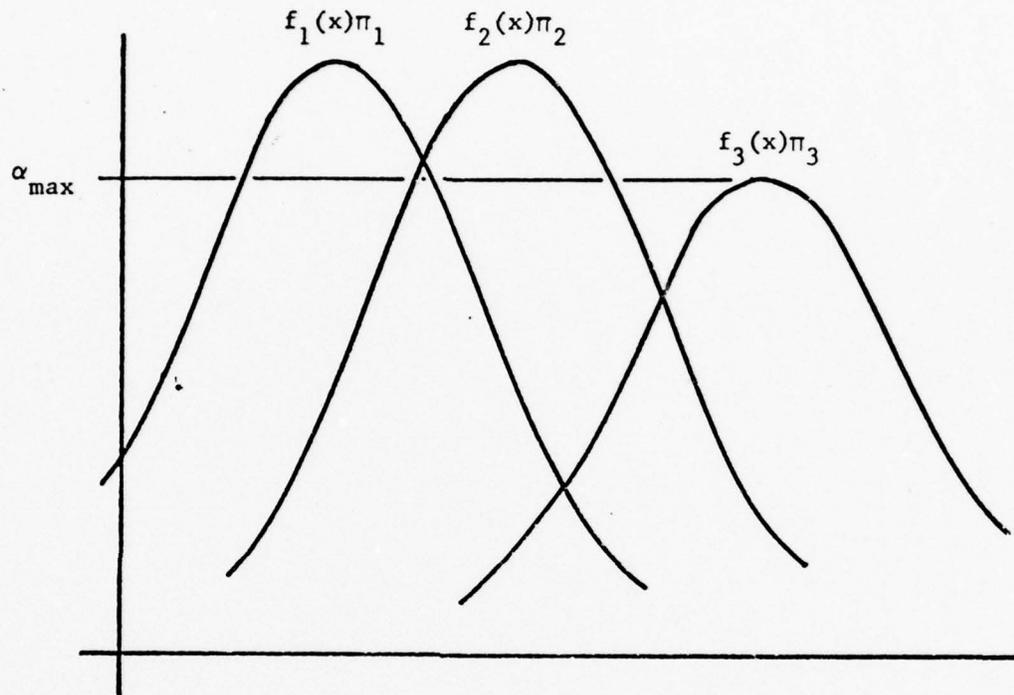


Figure 3-2. For all  $\alpha < \alpha_{max}$ ,  $1 \in T_2(\alpha)$  and  $2 \in T_1(\alpha)$ , thus the error count estimator is not in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{max}\}$  for these densities.

As  $\alpha$  varies between 0 and  $\alpha_{\max}$ , a family of unbiased, consistent estimators  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  is obtained. Each  $\alpha$  does not determine a unique form of an estimator since each  $\alpha$  does not uniquely determine the sets  $T_m(\alpha)$ ,  $m=1,2,\dots,M$ . In fact, the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  contains at most  $\frac{M(M-1)}{2} + 1$  different estimators. However, the value of the parameter  $\alpha$  is important in determining the density evaluations required by the estimator  $\hat{R}(\alpha)$ .

When  $\alpha = 0$ , the estimator  $\hat{R}(0)$  is equivalent to the posterior estimator  $\hat{R}(p)$ , of eq. (3-7), in the sense that estimates of the risk resulting from either estimator are identical and their variances are the same. However, as a member of the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  it is possible (if the conditional densities have finite support) that  $\hat{R}(0)$  may be computed with fewer density evaluations.

If  $\exists \alpha_e < \alpha_{\max}$  such that  $T_m(\alpha_e) = \{m\}$ ,  $m=1,2,\dots,M$  then  $\hat{R}(\alpha_e)$  is equivalent to the error count estimator  $\hat{R}(ec)$  of (3-1). Thus the posterior estimator is always equivalent to a member of the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  while the error count may or may not be. For the class densities in figure 3-2 the error count estimator would not be allowed in the family since  $\forall \alpha < \alpha_{\max}$ ,  $1 \in T_2(\alpha)$  and  $2 \in T_1(\alpha)$ .

#### 3.2.4 Computational Requirements for Estimators in The Family

The computational requirements for an estimator in the family will be given by the expected number of class conditional density evaluations it requires per sample. As in section 3.2.2, let us first consider the number of density evaluations explicitly required by the estimator  $\hat{R}(\alpha)$ , disregarding the fact that the error function  $\mathcal{E}_m(X_j)$  must be computed for each sample  $X_j$  and for each class  $m \in T_{\theta_j}(\alpha)$ . Then from equation

(3-29), at each point  $X_j$  the densities  $f_\ell(X_j)$  for all classes  $\ell \in T_m(\alpha)$ , for all  $m \in T_{\theta_j}(\alpha)$  must be computed. As in definition 3-1, let the sets  $Q_m(\alpha)$ ,  $m=1,2,\dots,M$  be

$$Q_m(\alpha) = \bigcup_{q \in T_m(\alpha)} T_q(\alpha) \quad (3-30)$$

Then for each sample  $X_j$ ,  $j=1,2,\dots,N$ ,  $\hat{R}(\alpha)$  requires explicitly the conditional densities  $f_\ell(X_j) \forall \ell \in Q_{\theta_j}(\alpha)$ .

Define the modified error function  $\mathcal{E}_m(X, Q_\theta(\alpha))$  for  $m \in T_\theta(\alpha)$  as

Definition 3-5

$$\mathcal{E}_m(X, Q_\theta(\alpha)) = \begin{cases} 0 & f_m(X)\pi_m > f_\ell(X)\pi_\ell \forall \ell \in Q_\theta(\alpha), \ell \neq m \\ 1 & \exists k \in Q_\theta(\alpha) \ni \\ & f_k(X)\pi_k > f_m(X)\pi_m \end{cases}$$

Then the modified error function may be evaluated on the basis of only those conditional densities  $f_\ell(X)$ ,  $\ell \in Q_\theta(\alpha)$  explicitly required at  $X$  by  $\hat{R}(\alpha)$ .

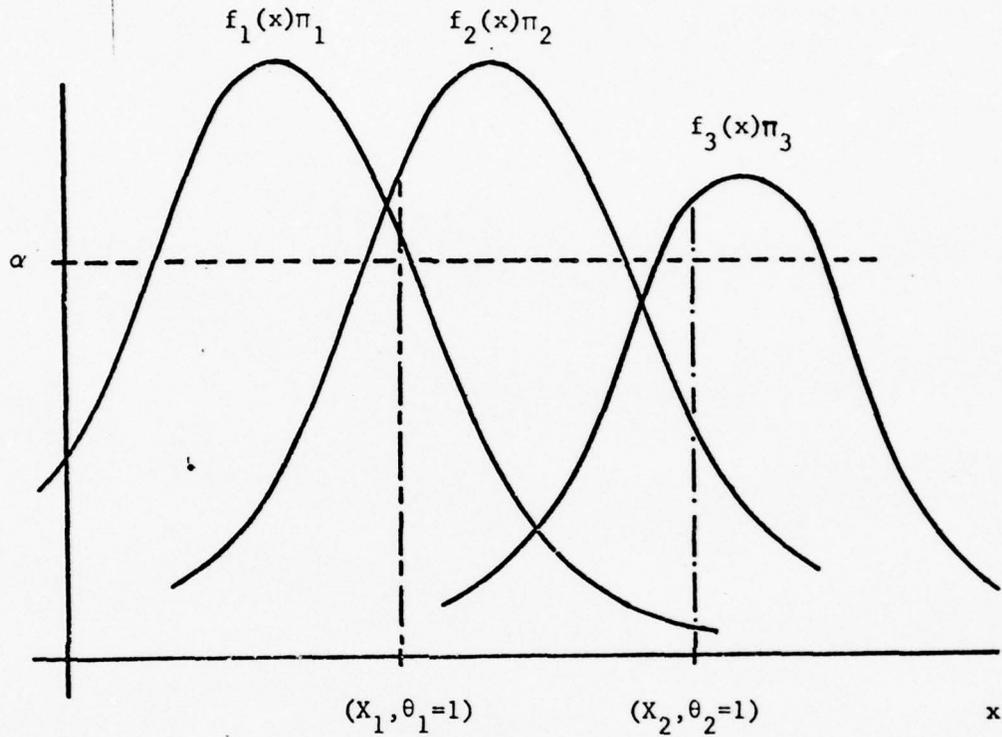
The following theorem shows that the modified error function is equal to the error function whenever the joint density of a sample  $X$  and its label  $\theta$  is greater than  $\alpha$ .

Theorem 3-2

For the random vector  $(X, \theta)$ , if  $m \in T_\theta(\alpha)$  and if  $f_\theta(X)\pi_\theta > \alpha$  then  $\mathcal{E}_m(X, Q_\theta(\alpha)) = \mathcal{E}_m(X)$ .

Proof

$\mathcal{E}_m(X, Q_\theta(\alpha)) = 1 \Rightarrow \mathcal{E}_m(X) = 1$  since if  $\exists k \in Q_\theta(\alpha) \ni f_k(X)\pi_k > f_m(X)\pi_m$  then  $\exists k \in \{1, 2, \dots, M\} \ni f_k(X)\pi_k > f_m(X)\pi_m$ . We will show that  $\mathcal{E}_m(X) = 1 \Rightarrow \mathcal{E}_m(X, Q_\theta(\alpha)) = 1$  by contradiction. For suppose  $\mathcal{E}_m(X) = 1$  and



$T_1(\alpha) = \{1, 2\}$	$Q_1(\alpha) = \{1, 2\}$
$T_2(\alpha) = \{1, 2\}$	$Q_2(\alpha) = \{1, 2\}$
$T_3(\alpha) = \{3\}$	$Q_3(\alpha) = \{3\}$

Figure 3-3. Modified error function equals true error function for sample  $X_1$  since  $f_{\theta_1}(X_1)\pi_{\theta_1} > \alpha$ , but since  $f_{\theta_2}(X_2)\pi_{\theta_2} \leq \alpha$  the true error function must be used for sample  $X_2$ .

$\mathcal{E}_m(X, Q_\theta(\alpha)) = 0$ . Then  $f_m(X)\pi_m > f_\ell(X)\pi_\ell \forall \ell \in Q_\theta(\alpha)$  but  $\exists k \notin Q_\theta(\alpha)$  such that  $f_k(X)\pi_k > f_m(X)\pi_m$ . But since  $\theta \in T_\theta(\alpha)$  and  $T_\theta(\alpha) \subset Q_\theta(\alpha)$ ,  $\theta \in Q_\theta(\alpha)$  so  $f_k(X)\pi_k > f_m(X)\pi_m \geq f_\theta(X)\pi_\theta > \alpha$ . Thus  $k \in T_\theta(\alpha)$  and since  $T_\theta(\alpha) \subset Q_\theta(\alpha)$ ,  $k \in Q_\theta(\alpha)$ , a contradiction.

Example 3-2 Consider the three class densities in figure 3-3 and the sets  $T_m(\alpha)$  and  $Q_m(\alpha)$  associated with the given  $\alpha$ . Since  $\theta_1 = \theta_2 = 1$  and  $T_1(\alpha) = \{1, 2\}$ , the functions  $\mathcal{E}_1(X_1)$ ,  $\mathcal{E}_2(X_1)$ ,  $\mathcal{E}_1(X_2)$  and  $\mathcal{E}_2(X_2)$  must be determined. Since  $f_1(X_1)\pi_1 > \alpha$ ,  $\mathcal{E}_1(X_1, Q_1(\alpha)) = \mathcal{E}_1(X_1) = 1$  and  $\mathcal{E}_2(X_1, Q_1(\alpha)) = \mathcal{E}_2(X_1) = 0$ . However, for the sample  $X_2$ ,  $\mathcal{E}_2(X_2, Q_1(\alpha)) = 0$  but  $\mathcal{E}_2(X_2) = 1$ . Since  $f_1(X_2)\pi_1 \leq \alpha$ , the conditional density  $f_3(X_2)$  must be computed.

Define the indicator function for the event  $f_\theta(X)\pi_\theta > \alpha$  as

Definition 3-6

$$I_\theta(X, \alpha) = \begin{cases} 1 & f_\theta(X)\pi_\theta > \alpha \\ 0 & f_\theta(X)\pi_\theta \leq \alpha \end{cases}$$

Then as a corollary to theorem 3-2, we have

Corollary 3-2 If  $m \in T_\theta(\alpha)$  then

$$I_\theta(X, \alpha) \mathcal{E}_m(X, Q_\theta(\alpha)) = I_\theta(X, \alpha) \mathcal{E}_m(X).$$

By corollary 3-2, a computational form for the estimator  $\hat{R}(\alpha)$  is given by

$$c\hat{R}(\alpha) = \frac{1}{N} \sum_{j=1}^N \sum_{m \in T_{\theta_j}(\alpha)} \{ I_{\theta_j}(X_j, \alpha) \mathcal{E}_m(X_j, Q_{\theta_j}(\alpha)) + (1 - I_{\theta_j}(X_j, \alpha)) \mathcal{E}_m(X_j) \} \frac{f_m(X_j)\pi_m}{\sum_{\ell \in T_m(\alpha)} f_\ell(X_j)\pi_\ell} \quad (3-31)$$

Note that while the estimator  $\hat{R}(\alpha)$  itself depends on  $\alpha$  only through the sets  $T_m(\alpha)$ ,  $m=1, 2, \dots, M$ , the computational form  $c\hat{R}(\alpha)$  uses  $\alpha$  directly

to determine which densities need to be evaluated. For if  $(X_j^{\theta_j})$  is such that  $f_{\theta_j}(X_j)\pi_{\theta_j} > \alpha$  then only those densities  $f_{\ell}(X_j)$  for  $\ell \in Q_{\theta_j}(\alpha)$  must be computed, since by theorem 3-2, the modified error function  $\mathcal{E}_m(X_j, Q_{\theta_j}(\alpha))$  is equal to the true error function  $\mathcal{E}_m(X_j)$ , for  $m \in T_{\theta_j}(\alpha)$ . If  $f_{\theta_j}(X_j)\pi_{\theta_j} \leq \alpha$  then all densities  $f_{\ell}(X_j)$   $\ell=1,2,\dots,M$  must be computed.

The total number of density evaluations required by  $\hat{cR}(\alpha)$  is thus

$$\sum_{j=1}^N I_{\theta_j}(X_j, \alpha) |Q_{\theta_j}(\alpha)| + (1 - I_{\theta_j}(X_j, \alpha))M, \quad (3-32)$$

where  $|Q_{\theta_j}(\alpha)|$  is the number of classes in the set  $Q_{\theta_j}(\alpha)$ . The expression above depends on the actual values of the sample  $\{(X_1^{\theta_1}), (X_2^{\theta_2}), \dots, (X_N^{\theta_N})\}$ . What we are really interested in is the expected number of density evaluations required in  $\hat{cR}(\alpha)$ , which is just the expectation of (3-32).

Let  $C(\alpha)$  be the expected number of density evaluations per sample required by  $\hat{cR}(\alpha)$ . Then

$$\begin{aligned} C(\alpha) &= E\{I_{\theta}(X, \alpha) |Q_{\theta}(\alpha)| + (1 - I_{\theta}(X, \alpha))M\} \\ &= \sum_{\ell=1}^M \int_S [I_{\ell}(x, \alpha) |Q_{\ell}(\alpha)| + (1 - I_{\ell}(x, \alpha))M] f_{\ell}(x)\pi_{\ell} dx \\ &= \sum_{\ell=1}^M [ |Q_{\ell}(\alpha)| \int_S I_{\ell}(x, \alpha) f_{\ell}(x)\pi_{\ell} dx \\ &\quad + M \int_S (1 - I_{\ell}(x, \alpha)) f_{\ell}(x)\pi_{\ell} dx ] . \end{aligned} \quad (3-33)$$

Let

$$\begin{aligned} U_{\ell}(\alpha) &= \int_S (1 - I_{\ell}(x, \alpha)) f_{\ell}(x)\pi_{\ell} dx \\ &= \int_{x \in S} f_{\ell}(x)\pi_{\ell} dx \\ &\quad \ni f_{\ell}(x)\pi_{\ell} \leq \alpha \end{aligned} \quad (3-34)$$

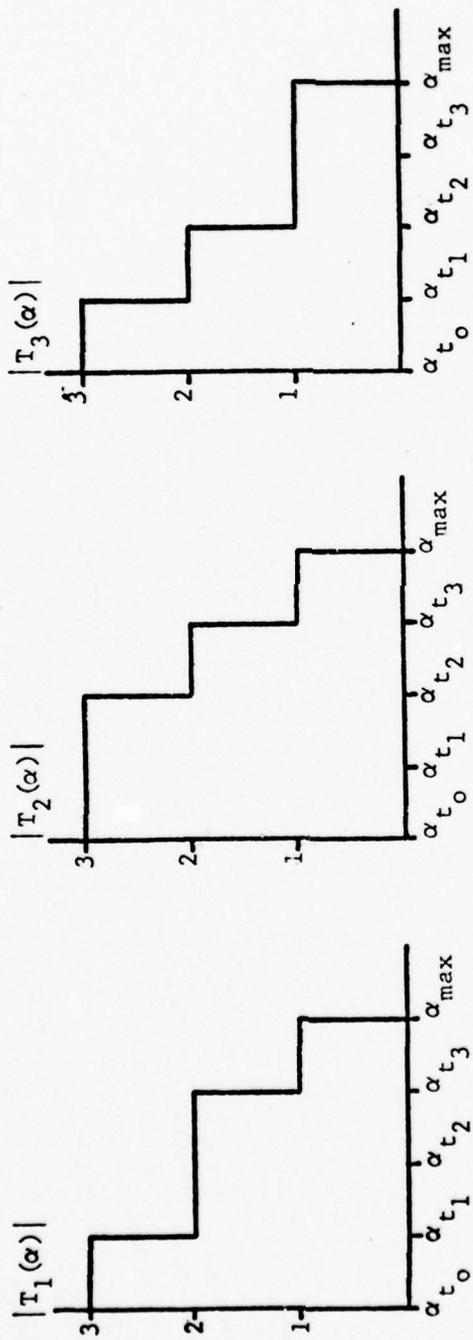


Figure 3-4. The number of elements  $|T_m(\alpha)|$  in  $T_m(\alpha)$  as a function of  $\alpha$ .

Then (3-33) may be written as

$$C(\alpha) = \sum_{\ell=1}^M [ |Q_{\ell}(\alpha)| (\pi_{\ell} - U_{\ell}(\alpha)) + MU_{\ell}(\alpha) ] \quad (3-35)$$

The total expected number of density evaluations required by  $\hat{cR}(\alpha)$  is just  $NXC(\alpha)$ , the expected number per sample times the number of samples.

Consider now the behavior of  $C(\alpha)$  as a function of  $\alpha$ . It is clear that for each  $m=1,2,\dots,M$ ,  $|T_m(\alpha)|$ , the number of classes in the set  $T_m(\alpha)$ , is non-increasing in  $\alpha$ . The points  $\alpha_{t_0}, \alpha_{t_1}, \dots, \alpha_{t_K}$  given by definition 3-4 are the values of  $\alpha$  which cause a decrease in  $|T_m(\alpha)|$  for some  $m=1,2,\dots,M$ . Figure 3-4 shows the behavior of  $|T_m(\alpha)|$ ,  $m=1,2,3$  for the class densities given in figure 3-1.

Recall that the sets  $Q_m(\alpha)$ ,  $m=1,2,\dots,M$ , were defined in terms of the sets  $T_m(\alpha)$ ,  $m=1,2,\dots,M$  as  $Q_m(\alpha) = \bigcup_{q \in T_m(\alpha)} q$ ,  $m=1,2,\dots,M$ . Thus  $|Q_m(\alpha)|$ , the number of classes in the set  $Q_m(\alpha)$ , is also non-increasing in  $\alpha$  for each  $m=1,2,\dots,M$ . Analogous to the points  $\alpha_{t_0}, \alpha_{t_1}, \dots, \alpha_{t_K}$ , we define the points  $\alpha_{q_0}, \alpha_{q_1}, \dots, \alpha_{q_J}$  that induce changes in the sets  $Q_m(\alpha)$  recursively as follows.

Definition 3-7

Let  $\alpha_{q_0} = 0$ .

Do  $i=0$  by 1 while  $\alpha_{q_i} < \alpha_{\max}$

Let  $\alpha_{q_{i+1}}$  be the smallest value of  $\alpha > \alpha_{q_i}$  such that

$Q_m(\alpha_{q_{i+1}}) \neq Q_m(\alpha_{q_i})$  for some  $m=1,2,\dots,M$ .

End.

Let  $\alpha_{q_J}$  be the largest value so defined.

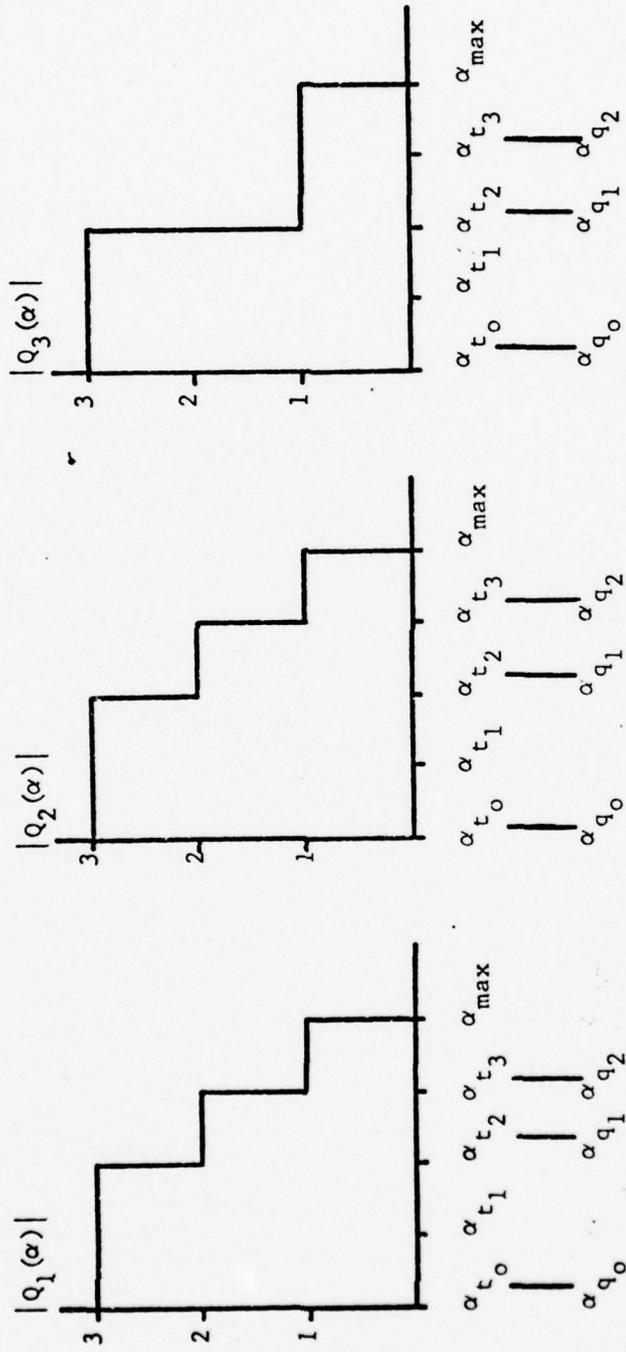


Figure 3-5. The number of elements  $|Q_m(\alpha)|$  in  $Q_m(\alpha)$  as a function of  $\alpha$ .

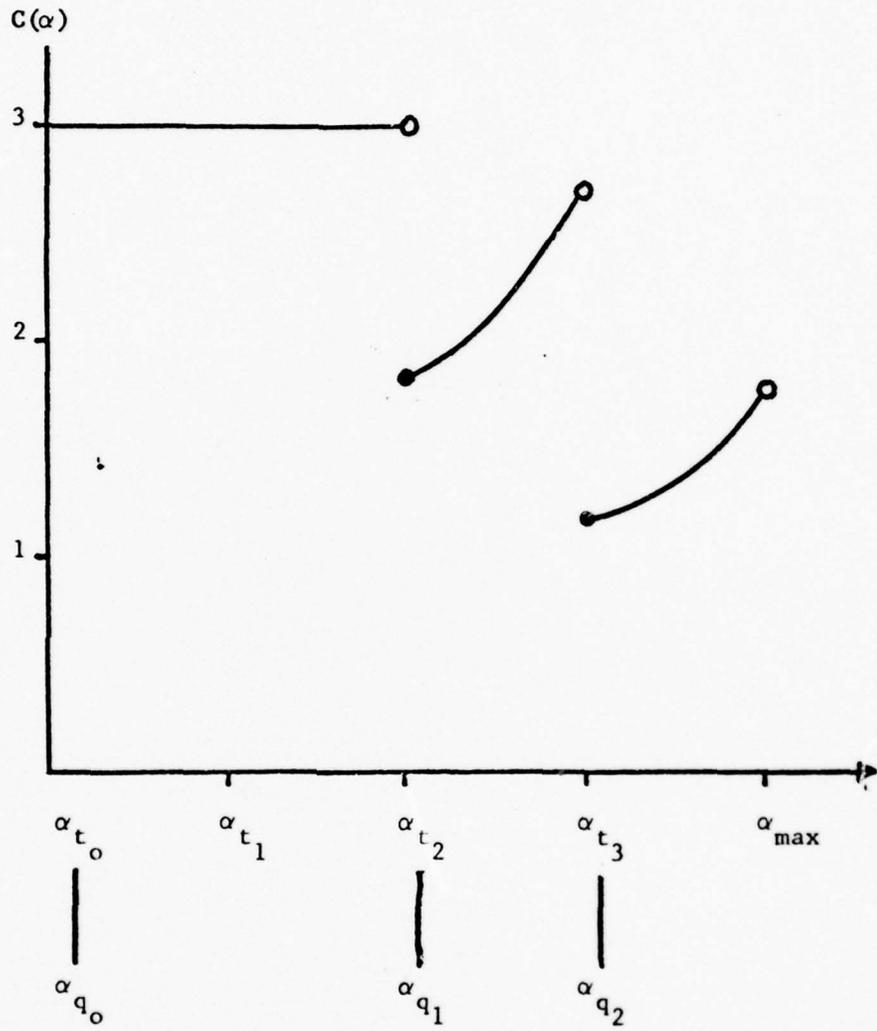


Figure 3-6. The expected number of density evaluations per sample  $C(\alpha)$  as a function of  $\alpha$ .

Then the values  $\alpha_{q_0}, \alpha_{q_1}, \dots, \alpha_{q_J}$  are those values of  $\alpha$  which cause a decrease in  $|Q_m(\alpha)|$  for some  $m$ . Also, it is clear that for each  $j=0,1,\dots,J$   $\alpha_{q_j} = \alpha_{t_i}$  for some  $i=0,1,\dots,K$ . In figure 3-1, the points  $\alpha_{q_0}, \alpha_{q_1}$  and  $\alpha_{q_2}$  are given, and Table 3-2 shows the sets  $Q_m(\alpha)$   $m=1,2,3$ . Figure 3-5 shows the behavior of  $|Q_m(\alpha)|$   $m=1,2,3$  for the class densities in figure 3-1.

From (3-34), it is clear that  $U_\ell(\alpha)$ ,  $\ell=1,2,\dots,M$  is a non-decreasing function of  $\alpha$ . Rewriting the expression (3-35) for  $C(\alpha)$  as

$$C(\alpha) = \sum_{\ell=1}^M U_\ell(\alpha) (M - |Q_\ell(\alpha)|) + \pi_\ell |Q_\ell(\alpha)| \quad (3-36)$$

we see that for  $\alpha_{q_i} \leq \alpha < \alpha_{q_{i+1}}$ ,  $C(\alpha)$  is non-decreasing, since  $\forall \ell=1,2,\dots,M$ ,  $U_\ell(\alpha)$  is a non-decreasing and  $|Q_\ell(\alpha)|$  is constant. Figure 3-6 shows a schematic drawing of  $C(\alpha)$  as a function of  $\alpha$  for the densities in figure 3-1.

### 3.2.5 Variances of Estimators in The Family

Next consider the variances of estimators in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$ . A given  $\alpha$  defines the sets  $T_m(\alpha)$ ,  $m=1,2,\dots,M$  and hence an estimator  $\hat{R}(\alpha)$  of the general form  $\hat{R}(T)$ . From the variance of  $\hat{R}(T)$  given by (3-19), we may write

$$\text{VAR}\{\hat{R}(\alpha)\} = \frac{1}{N} \text{VAR}\left\{ \sum_{m \in T_\theta(\alpha)} \mathcal{E}_m(X) \frac{f_m(X)\pi_m}{\sum_{\ell \in T_m(\alpha)} f_\ell(X)\pi_\ell} \right\} \quad (3-37)$$

$$= \frac{1}{N} \left\{ \sum_{i=1}^M \pi_i \int_S \left( \sum_{m \in T_i(\alpha)} \mathcal{E}_m(x) \frac{f_m(x)\pi_m}{\sum_{\ell \in T_m(\alpha)} f_\ell(x)\pi_\ell} \right)^2 f_i(x) dx - R^2 \right\}.$$

Let the coefficient of variance  $V(\alpha)$  be defined by

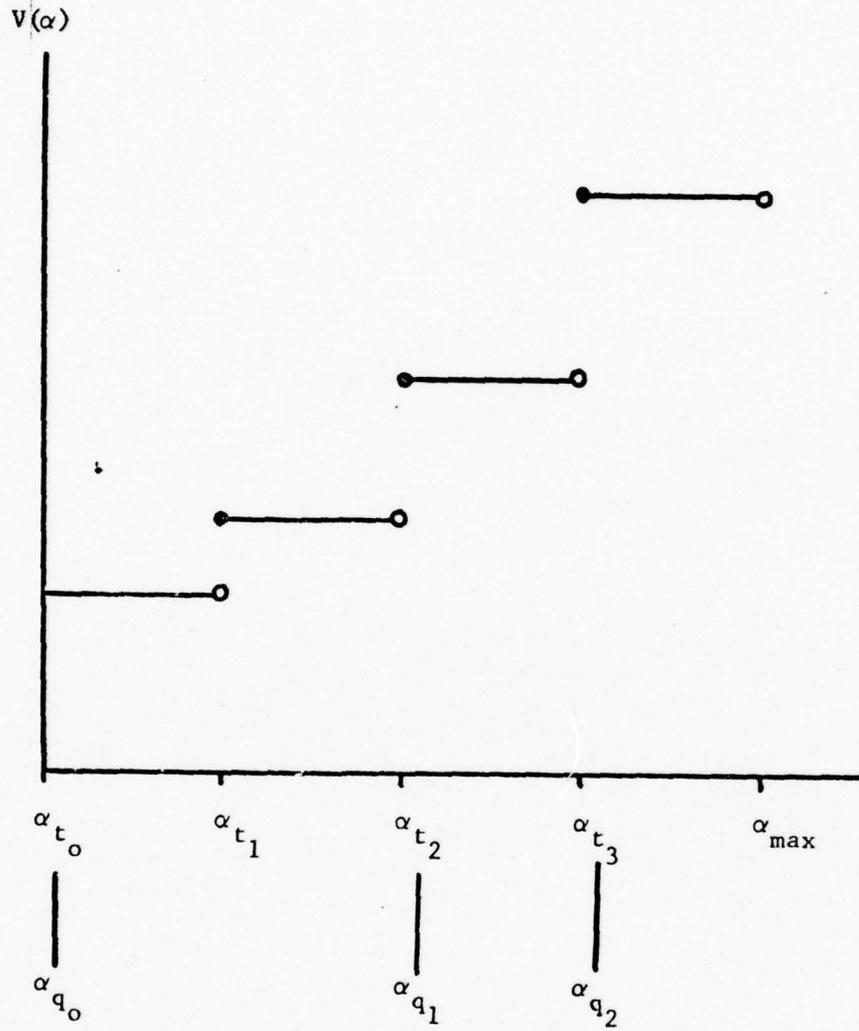


Figure 3-7. The coefficient of variance  $V(\alpha)$  as a function of  $\alpha$ .

$$V(\alpha) = N \times \text{VAR}\{\hat{R}(\alpha)\} \quad (3-38)$$

When  $\alpha=0$ ,  $\hat{R}(0)$  is equivalent to the posterior estimator  $\hat{R}(p)$ .

Theorem 3-1 shows that the variance of the posterior estimator is smaller than the variance of any estimator expressed in the general form, thus as a corollary we have,

Corollary 3-1  $V(0) \leq V(\alpha) \quad \forall 0 \leq \alpha < \alpha_{\max}$ .

If there exists an  $\alpha_e < \alpha_{\max}$  such that  $T_m(\alpha_e) = \{m\} \quad \forall m=1,2,\dots,M$ , then  $\hat{R}(\alpha_e)$  is equivalent to the error count estimator. Thus by (3-3),  $V(\alpha_e) = R(1-R)$ . By corollary 3-1,  $V(0) \leq V(\alpha_e)$ .

Now  $V(\alpha)$  depends on  $\alpha$  only through the sets  $T_m(\alpha)$ ,  $m=1,2,\dots,M$  (see (3-37)). Since the points  $\alpha_{t_0}, \alpha_{t_1}, \dots, \alpha_{t_K}$  induce changes in these sets,

$V(\alpha)$  is a step function of  $\alpha$  with discontinuities at these points.

Figure 3-7 gives a schematic drawing of  $V(\alpha)$  for the densities in figure 3-1.

Examples indicate that  $V(\alpha)$  is a non-decreasing function of  $\alpha$ .

Consider the estimator  $\hat{R}_m(\alpha)$  for the conditional risk  $\hat{R}_m$  given in (3-28).

The variance of  $\hat{R}_m(\alpha)$  is

$$\text{VAR}\{\hat{R}_m(\alpha)\} = \frac{1}{N} \left\{ \int_S \mathcal{E}_m(x) \frac{f_m^2(x)}{\sum_{q \in T_m(\alpha)} f_q(x) \pi_q} dx - R_m^2 \right\}. \quad (3-39)$$

Define the coefficient of variance  $V_m(\alpha)$  for  $\hat{R}_m(\alpha)$  as

$$V_m(\alpha) = N \times \text{VAR}\{\hat{R}_m(\alpha)\}. \quad (3-40)$$

Then  $V_m(\alpha)$  is a non-decreasing function of  $\alpha$  since, as  $\alpha$  increases, the number of classes in  $T_m(\alpha)$  is non-increasing and hence

$$\sum_{q \in T_m(\alpha)} f_q(x) \pi_q$$

is non-increasing.

The covariance of the estimators  $\hat{R}_m(\alpha)$  and  $\hat{R}_l(\alpha)$  is given by

$$\text{COV}\{\hat{R}_m(\alpha), \hat{R}_l(\alpha)\} = \quad (3-41)$$

$$\frac{1}{N} \left\{ \sum_{i \in T_m(\alpha) \cap T_l(\alpha)} \pi_i \int_S e_m(x) e_l(x) \frac{f_m(x) f_l(x) f_i(x)}{\sum_{q \in T_m(\alpha)} f_q(x) \pi_q \sum_{r \in T_l(\alpha)} f_r(x) \pi_r} dx - R_m R_l \right\}$$

where  $\sum_{i \in T_m(\alpha) \cap T_l(\alpha)} (\quad) = 0$

when the intersection of  $T_m(\alpha)$  and  $T_l(\alpha)$  is empty, i.e.  $T_m(\alpha) \cap T_l(\alpha) = \emptyset$

Let  $C_{ml}(\alpha) = N \times \text{COV}\{\hat{R}_m(\alpha), \hat{R}_l(\alpha)\}$  (3-42)

be the coefficient of covariance. Since  $\hat{R}(\alpha) = \sum_{m=1}^M \pi_m \hat{R}_m(\alpha)$ , the coefficient of variance  $V(\alpha)$  for  $\hat{R}(\alpha)$  may be expressed in terms of  $V_m(\alpha)$  and  $C_{ml}(\alpha)$  as

$$V(\alpha) = \sum_{m=1}^M \pi_m^2 V_m(\alpha) + \sum_{m=1}^M \sum_{l \neq m} \pi_m \pi_l C_{ml}(\alpha) \quad (3-43)$$

Now  $\forall m=1, 2, \dots, M$ ,  $V_m(\alpha)$  is non-decreasing in  $\alpha$  and  $V_m(0) \leq V_m(\alpha)$ ,

$\forall 0 \leq \alpha < \alpha_{\max}$ . However, from (3-41),  $C_{ml}(\alpha)$  achieves its minimum value

of  $-R_m R_l$  when  $T_m(\alpha) \cap T_l(\alpha) = \emptyset$ , which would occur for large values of  $\alpha$ .

Thus if the increase in the conditional variances  $V_m(\alpha)$  dominate the possible decreases in  $C_{ml}(\alpha)$ , as  $\alpha$  increases one would expect the co-

efficient of variance  $V(\alpha)$  to increase.

### 3.2.6 Examples

Some examples are given in the appendix. In example 1, there are five equally likely classes. The class conditional densities are Gaussian with standard deviations equal to one and means 0, .75, 7, 8, 9.5.

Page A-1 gives a sketch of the densities as well as the true conditional

and unconditional risks. Page A-2 gives the values of  $\alpha_{t_0}, \alpha_{t_1} \dots \alpha_{t_{10}}$ ,  $\alpha_{q_0}, \alpha_{q_1} \dots \alpha_{q_5}$  and the sets  $T_m(\alpha_{t_i}), Q_m(\alpha_{t_i}), m=1,2,\dots,5$  which they determine. Page A-3 gives, for the points  $\alpha_{t_i}$ , the expected number of density evaluations  $C(\alpha)$  and the coefficient of variance  $V(\alpha)$  for those densities. Note that  $C(\alpha)$  decreases to a minimum of 2.6 at  $\alpha_{t_6}$ . The sets  $T_m(\alpha_{t_6}), m=1,2,\dots,5$  on page A-2 are seen to be the natural grouping of the classes. The increase in  $C(\alpha)$  for  $\alpha > \alpha_{t_6}$  is due to the fact that while the sets  $T_m(\alpha)$  and  $Q_m(\alpha)$  are getting smaller, for larger  $\alpha$  it is becoming less likely that a sample  $X$  has the property that  $f_\theta(X)\pi_\theta > \alpha$ . When  $f_\theta(X)\pi_\theta \leq \alpha$ , all densities  $f_\ell(X), \ell=1,2,\dots,5$  must be computed to determine the error function, so the smallness of the sets  $Q_m(\alpha)$  becomes irrelevant.

The variances  $V(\alpha)$  are clearly non-decreasing in  $\alpha$ . For  $\alpha \leq \alpha_{t_6}$ , they are approximately the same. For  $\alpha > \alpha_{t_6}$ , the variances about double at each decrease in some set  $T_m$ . Page A-4 shows the covariance matrix for the conditional risk estimators  $\hat{R}_m(\alpha), \hat{R}_\ell(\alpha), m,\ell=1,2,\dots,M$  for  $\alpha = \alpha_{t_0}, \alpha_{t_8}, \alpha_{t_{10}}$ . Note that the increase in variance as  $\alpha$  increases seems to be due mostly to the variance increase in  $\hat{R}_m(\alpha)$ , rather than to changes in the covariance of  $\hat{R}_m(\alpha)$  and  $\hat{R}_\ell(\alpha)$ .

Page A-5 shows the behavior of the distinct estimators  $\hat{R}(\alpha_{t_i})$  for various sample sizes where larger variances for larger  $\alpha$ 's are reflected. For  $\alpha = \alpha_{t_0}$ , the estimator  $\hat{R}(\alpha_{t_0})$  is equivalent to the posterior estimator  $\hat{R}(p)$ . For  $\alpha = \alpha_{t_{10}}$ ,  $\hat{R}(\alpha_{t_{10}})$  is equivalent to the error count estimator.

In example 2, the five classes have unequal priors given by .1, .3,

.2, .19, .21. The class conditional densities are normal with means 0, .5, 6, 10, 11 and equal standard deviations of 1. The densities are sketched on page B-1 and the true conditional and unconditional risks are given. Page B-2 gives the values  $\alpha_{t_0}, \alpha_{t_1}, \dots, \alpha_{t_8}, \alpha_{q_0}, \dots, \alpha_{q_3}$  and the resulting sets  $T_m$  and  $Q_m$ . Note that the error count estimator is not included in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$ .

Page B-3 gives  $C(\alpha)$  and  $V(\alpha)$  for the points  $\alpha_{t_i}, i=0, \dots, 8$ . The expected number of density evaluations per sample  $C(\alpha)$  decreases to a minimum of 1.94 at  $\alpha_{t_8}$ . Again the variance  $V(\alpha)$  appears to be increasing in  $\alpha$ .

Page B-4 shows the behavior of  $\hat{R}(\alpha)$  for  $\alpha = \alpha_{t_i}, i=0, \dots, 8$ .  $\hat{R}(\alpha_{t_0})$  is equivalent to the posterior estimator. The error count estimator is included for reference.

In chapter 4, the problem of choosing an optimal estimator from the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  is discussed. Considerations of optimality will involve the variance  $V(\alpha)$  of the estimators and  $C(\alpha)$ , the expected number of density evaluations per sample.

But first, the technique of stratified sampling will be discussed in terms of a family of risk estimators.

### 3.3 Estimators Based on Stratified Sampling

Stratified sampling is a classic Monte Carlo technique for reducing the variance of an estimator for an integral [31, 19, 39, 16]. Basically, one partitions the region of integration and samples independently from each partition. In this case, the integral to be estimated is

$$R = \sum_{m=1}^M \pi_m \int_S e_m(x) f_m(x) dx.$$

The summation represents a natural partition of the integral. Thus rather than sampling  $(X, \theta)$  where  $\theta$  is random, the class  $\theta=m$ , is fixed a priori and observations are sampled independently from the distribution of  $X$  given  $m$ , for  $m=1, 2, \dots, M$ .

Let the stratified sample of size  $N$  be denoted by  $\{X_{11}, X_{12}, \dots, X_{1n_1}\}, (X_{21}, X_{22}, \dots, X_{2n_2}), \dots, (X_{M1}, X_{M2}, \dots, X_{Mn_M})\}$  where  $N = \sum_{i=1}^M n_i$ . The samples  $X_{ij}$ ,  $i=1, 2, \dots, M$ ,  $j=1, 2, \dots, n_i$  are independent, and the distribution of  $X_{ij}$ ,  $j=1, 2, \dots, n_i$  is identical to that of  $X_i$ . The density of  $X_i$  is given by  $f_i(x)$ , the conditional density of  $X$  given class  $i$ . The statistician is free to choose the number  $n_i$  of samples from class  $i$ ,  $i=1, 2, \dots, M$  in any way such that  $\sum_{i=1}^M n_i = N$ . Optimal and heuristic choices of these samples sizes will be discussed in section 3.3.2.

### 3.3.1 A Parameterized Family of Bayes Risk Estimators

A family of estimators for the Bayes risk  $R$  based on stratified sampling is defined as  $\{\hat{SR}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$ . The scalar parameter  $\alpha$  determines the set of classes  $T_m(\alpha)$  that are " $\alpha$ -close" to class  $m$ ,  $m=1, 2, \dots, M$  as in section 3.2.3. The stratified estimator  $\hat{SR}_m(\alpha)$  for the conditional risk  $R_m$  is based on samples  $X_{ij}$  for  $i \in T_m(\alpha)$ ,  $j=1, 2, \dots, n_i$  and is given by

$$\hat{SR}_m(\alpha) = \sum_{i \in T_m(\alpha)} \frac{\pi_i}{n_i} \sum_{j=1}^{n_i} e_m(X_{ij}) \frac{f_m(X_{ij})}{\sum_{l \in T_m(\alpha)} f_l(X_{ij})} \quad (3-44)$$

$\hat{SR}_m(\alpha)$  is an unbiased estimator for  $R_m$  since

$$E\{\hat{SR}_m(\alpha)\} = \sum_{i \in T_m(\alpha)} \pi_i E\{e_m(X_i)\} \frac{f_m(X_i)}{\sum_{l \in T_m(\alpha)} f_l(X_i) \pi_l}$$

$$= \sum_{i \in T_m(\alpha)} \pi_i \int_S \mathcal{E}_m(x) \frac{f_m(x) f_i(x) dx}{\sum_{l \in T_m(\alpha)} f_l(x) \pi_l} = \int_S \mathcal{E}_m(x) f_m(x) dx = R_m. \quad (3-45)$$

The estimator  $\hat{SR}(\alpha)$  for the unconditional risk  $R$  is defined

$$\begin{aligned} \hat{SR}(\alpha) &= \sum_{m=1}^M \pi_m \hat{SR}_m(\alpha) \\ &= \sum_{m=1}^M \pi_m \sum_{i \in T_m(\alpha)} \frac{\pi_i}{n_i} \sum_{j=1}^{n_i} \mathcal{E}_m(X_{ij}) \frac{f_m(X_{ij})}{\sum_{l \in T_m(\alpha)} f_l(X_{ij}) \pi_l} \end{aligned} \quad (3-46)$$

and is unbiased by linearity of the expectation operator. Also, since

$i \in T_m(\alpha)$  iff  $m \in T_i(\alpha)$ ,  $\sum_{m=1}^M \sum_{i \in T_m(\alpha)} = \sum_{i=1}^M \sum_{m \in T_i(\alpha)}$ , and thus  $\hat{SR}(\alpha)$  may be written

$$\hat{SR}(\alpha) = \sum_{i=1}^M \frac{\pi_i}{n_i} \sum_{j=1}^{n_i} \sum_{m \in T_i(\alpha)} \mathcal{E}_m(X_{ij}) \frac{f_m(X_{ij}) \pi_m}{\sum_{l \in T_m(\alpha)} f_l(X_{ij}) \pi_l}. \quad (3-47)$$

When  $\alpha=0$ ,  $\hat{SR}(0)$  is equivalent to the stratified posterior estimator  $\hat{SR}(p)$ , where  $\hat{SR}(p)$  is given as in [39] by

$$\begin{aligned} \hat{SR}(p) &= \sum_{i=1}^M \frac{\pi_i}{n_i} \sum_{j=1}^{n_i} r(X_{ij}) \\ &= \sum_{i=1}^M \frac{\pi_i}{n_i} \sum_{j=1}^{n_i} \sum_{m=1}^M \mathcal{E}_m(X_{ij}) \frac{f_m(X_{ij}) \pi_m}{\sum_{l=1}^M f_l(X_{ij}) \pi_l} \end{aligned} \quad (3-48)$$

Also, if  $\exists \alpha_e < \alpha_{\max}$  such that  $T_m(\alpha_e) = \{m\} \forall m=1, 2, \dots, M$  then  $\hat{SR}(\alpha_e)$  is equivalent to the stratified error count estimator  $\hat{SR}(ec)$ , where  $\hat{SR}(ec)$  is [39],

$$\hat{SR}(ec) = \sum_{i=1}^M \frac{\pi_i}{n_i} \sum_{j=1}^{n_i} \mathcal{E}_i(X_{ij}). \quad (3-49)$$

3.3.2 Variations of Estimators in The Family

The variance of the estimator  $\hat{SR}(\alpha)$

$$\begin{aligned} \text{VAR}\{\hat{SR}(\alpha)\} &= \sum_{i=1}^M \frac{\pi_i^2}{n_i} \text{VAR}\left\{ \sum_{m \in T_i(\alpha)} \mathcal{E}_m(X_i) \frac{f_m(X_i)\pi_m}{\sum_{\ell \in T_m(\alpha)} f_\ell(X_i)\pi_\ell} \right\} \\ &= \sum_{i=1}^M \frac{\pi_i^2}{n_i} \left\{ \int_S \left( \sum_{m \in T_i(\alpha)} \mathcal{E}_m(x) \frac{f_m(x)\pi_m}{\sum_{\ell \in T_m(\alpha)} f_\ell(x)\pi_\ell} \right)^2 f_i(x) dx \right. \\ &\quad \left. - \left[ \int_S \sum_{m \in T_i(\alpha)} \mathcal{E}_m(x) \frac{f_m(x)\pi_m f_i(x) dx}{\sum_{\ell \in T_m(\alpha)} f_\ell(x)\pi_\ell} \right]^2 \right\}. \end{aligned} \quad (3-50)$$

A heuristic choice of the number of samples  $n_i$  from class  $i$  is to let  $n_i$  be proportional to the prior probability  $\pi_i$  of class  $i$ , thus  $n_i = N\pi_i$ . Even though this choice is not optimal, the estimator  $\hat{SR}(\alpha)$  based on the stratified sample has smaller variance than the estimator  $\hat{R}(\alpha)$  based on unrestricted sampling. To see this, with  $n_i = N\pi_i$ , the variance of  $\hat{SR}(\alpha)$  is given by

$$\begin{aligned} \text{VAR}\{\hat{SR}(\alpha)\} &= \\ &= \frac{1}{N} \left\{ \sum_{i=1}^M \pi_i \int_S \left( \sum_{m \in T_i(\alpha)} \mathcal{E}_m(x) \frac{f_m(x)\pi_m}{\sum_{\ell \in T_m(\alpha)} f_\ell(x)\pi_\ell} \right)^2 f_i(x) dx \right. \\ &\quad \left. - \sum_{i=1}^M \pi_i \left[ \int_S \sum_{m \in T_i(\alpha)} \mathcal{E}_m(x) \frac{f_m(x)\pi_m f_i(x) dx}{\sum_{\ell \in T_m(\alpha)} f_\ell(x)\pi_\ell} \right]^2 \right\}. \end{aligned} \quad (3-51)$$

By Jensen's inequality [7], and the fact that  $\sum_{i=1}^M \sum_{m \in T_i(\alpha)} = \sum_{m=1}^M \sum_{i \in T_m(\alpha)}$ ,

we have that

$$\begin{aligned}
& \sum_{i=1}^M \pi_i \left[ \int_S \sum_{m \in T_i(\alpha)} \mathcal{E}_m(x) \frac{f_m(x) \pi_m f_i(x)}{\sum_{l \in T_m(\alpha)} f_l(x) \pi_l} \right]^2 \\
& \geq \left[ \sum_{i=1}^M \pi_i \int_S \sum_{m \in T_i(\alpha)} \mathcal{E}_m(x) \frac{f_m(x) \pi_m f_i(x)}{\sum_{l \in T_m(\alpha)} f_l(x) \pi_l} dx \right]^2 \\
& = \left[ \sum_{m=1}^M \int_S \mathcal{E}_m(x) f_m(x) \pi_m \frac{\sum_{i \in T_m(\alpha)} f_i(x) \pi_i}{\sum_{l \in T_m(\alpha)} f_l(x) \pi_l} dx \right]^2 = R^2.
\end{aligned} \tag{3-52}$$

Thus,

$$\begin{aligned}
& \text{VAR}\{\hat{SR}(\alpha)\} \leq \\
& \frac{1}{N} \left\{ \sum_{i=1}^M \pi_i \int_S \left( \sum_{m \in T_i(\alpha)} \mathcal{E}_m(x) \frac{f_m(x) \pi_m}{\sum_{l \in T_m(\alpha)} f_l(x) \pi_l} \right)^2 f_i(x) dx - R^2 \right\} \\
& = \text{VAR}\{\hat{R}(\alpha)\}.
\end{aligned} \tag{3-53}$$

An optimal choice [39, 31] of the number of samples from each class is to choose  $n_i$ ,  $i=1,2,\dots,M$  to minimize the variance of the estimator  $\hat{SR}(\alpha)$ . Let

$$\begin{aligned}
\sigma_i^2(\alpha) &= \int_S \left( \sum_{m \in T_i(\alpha)} \mathcal{E}_m(x) \frac{f_m(x) \pi_m}{\sum_{l \in T_m(\alpha)} f_l(x) \pi_l} \right)^2 f_i(x) dx \\
&= \left[ \int_S \sum_{m \in T_i(\alpha)} \mathcal{E}_m(x) \frac{f_m(x) \pi_m f_i(x) dx}{\sum_{l \in T_m(\alpha)} f_l(x) \pi_l} \right]^2.
\end{aligned} \tag{3-54}$$

Then

$$\text{VAR}\{\hat{SR}(\alpha)\} = \sum_{i=1}^M \pi_i^2 \frac{\sigma_i^2(\alpha)}{n_i} . \quad (3-55)$$

For a given  $\alpha$ , the optimal choice of  $n_i$ ,  $i=1,2,\dots,M$  is found by solving the constrained minimization problem (N1);

$$\text{minimize } \sum_{i=1}^M \pi_i^2 \frac{\sigma_i^2(\alpha)}{n_i}$$

(N1)

$$\text{subject to } \sum_{i=1}^M n_i = N .$$

It is shown in [39, 31] that the solution to (N1) is

$$n_i^* = N \frac{\pi_i \sigma_i(\alpha)}{\sum_{\ell=1}^M \pi_{\ell} \sigma_{\ell}(\alpha)}, \quad i=1,2,\dots,M . \quad (3-56)$$

Thus the optimal choice of the sample sizes  $n_i = n_i^*$ ,  $i=1,2,\dots,M$  agrees with the heuristic choice  $n_i = N\pi_i$ ,  $i=1,2,\dots,M$  only when  $\sigma_i(\alpha) = \sigma(\alpha) \forall i=1,2,\dots,M$ . Of course, the problem with using the optimal choice is that knowledge of  $\sigma_i(\alpha)$ ,  $i=1,2,\dots,M$  is assumed, and if we knew this we would probably know the true risk  $R$ . Since choosing  $n_i$ ,  $i=1,2,\dots,M$  proportional to the prior probability of class  $i$  causes the stratified estimator  $\hat{SR}(\alpha)$  to have smaller variance than the unrestricted estimator  $\hat{R}(\alpha)$  anyway, we will assume in the sequel this heuristic choice of sample sizes.

As in unrestricted sampling, examples indicate that the variance of  $\hat{SR}(\alpha)$  is non-decreasing in  $\alpha$ . Moore, Whitsitt and Landgrebe [30] give a 2-class example where the stratified posterior estimator has smaller variance than the stratified error count estimator. However, for

for their example, the error count estimator would not be included in the family  $\{\hat{SR}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$ .

In fact, it is not clear (as it was with unrestricted sampling) that the variance of the estimator  $\hat{SR}_m(\alpha)$  of the conditional risk  $R_m$  is non-decreasing in  $\alpha$ . For in this case

$$\begin{aligned} \text{VAR}\{\hat{SR}_m(\alpha)\} &= \frac{1}{N} \left\{ \int_S \mathcal{E}_m(x) \frac{f_m^2(x)}{\sum_{q \in T_m(\alpha)} f_q(x) \pi_q} dx \right. \\ &\quad \left. - \sum_{i \in T_m(\alpha)} \pi_i \left[ \int_S \mathcal{E}_m(x) \frac{f_m(x) f_i(x) dx}{\sum_{q \in T_m(\alpha)} f_q(x) \pi_q} \right]^2 \right\}. \end{aligned} \quad (3-57)$$

For completeness, the covariance of the conditional stratified estimators  $\{\hat{SR}_m(\alpha) \text{ and } \hat{SR}_l(\alpha)\}$  is

$$\text{COV}\{\hat{SR}_m(\alpha), \hat{SR}_l(\alpha)\} = \quad (3-58)$$

$$\begin{aligned} &\frac{1}{N} \left\{ \sum_{i \in T_m(\alpha) \cap T_l(\alpha)} \pi_i \left[ \int_S \mathcal{E}_m(x) \mathcal{E}_l(x) \frac{f_m(x) f_l(x) f_i(x) dx}{\sum_{q \in T_m(\alpha)} f_q(x) \pi_q \sum_{r \in T_l(\alpha)} f_r(x) \pi_r} \right. \right. \\ &\quad \left. \left. - \int_S \mathcal{E}_m(x) \frac{f_m(x) f_i(x) dx}{\sum_{q \in T_m(\alpha)} f_q(x) \pi_q} \int_S \mathcal{E}_l(x) \frac{f_l(x) f_i(x) dx}{\sum_{r \in T_l(\alpha)} f_r(x) \pi_r} \right] \right\} \end{aligned}$$

which is zero when  $T_m(\alpha) \cap T_l(\alpha) = \emptyset$ .

### 3.3.3 Computational Requirements for Estimators in The Family

As for unrestricted sampling, a computational form for the stratified estimator  $\hat{SR}(\alpha)$  is defined

$$\begin{aligned}
\widehat{cSR}(\alpha) = & \sum_{i=1}^M \frac{\pi_i}{n_i} \sum_{j=1}^{n_i} \sum_{m \in T_i(\alpha)} (I_i(X_{ij}, \alpha) e_m(X_{ij}, Q_i(\alpha))) \\
& + (1 - I_i(X_{ij}, \alpha)) e_m(X_{ij}) \frac{f_m(X_{ij}) \pi_m}{\sum_{\ell \in T_m(\alpha)} f_\ell(X_{ij}) \pi_\ell}
\end{aligned} \tag{3-59}$$

$$\text{where } I_i(X_{ij}, \alpha) = \begin{cases} 1 & f_i(X_{ij}) \pi_i > \alpha, \\ 0 & f_i(X_{ij}) \pi_i \leq \alpha \end{cases}$$

The expected number of density evaluations per sample required in  $\widehat{cSR}(\alpha)$  is  $SC(\alpha)$ , where

$$SC(\alpha) = \frac{1}{N} \sum_{i=1}^M \frac{n_i}{\pi_i} (U_i(\alpha) (M - |Q_i(\alpha)|) + \pi_i |Q_i(\alpha)|). \tag{3-60}$$

With the heuristic choice  $n_i = N\pi_i$ ,  $i=1, 2, \dots, M$ , we have that

$$SC(\alpha) = \sum_{i=1}^M U_i(\alpha) (M - |Q_i(\alpha)|) + \pi_i |Q_i(\alpha)|. \tag{3-61}$$

Thus  $SC(\alpha)$  is equal to  $C(\alpha)$  (see 3-36), so that the expected number of density evaluations required by the stratified estimator  $\widehat{cSR}(\alpha)$  is the same as the number required by the unrestricted estimator  $\widehat{cR}(\alpha)$ .

## CHAPTER 4

## OPTIMAL ESTIMATORS

4.1 Introduction

Two families of unbiased, consistent estimators for the Bayes risk have been proposed:  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  for unrestricted sampling and  $\{\hat{SR}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  for stratified sampling. Given a sampling technique, the problem now is to choose an estimator in that family which is optimal for our purpose. We will restrict attention to the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$ . Extension to stratified sampling is obvious.

There are two major considerations in the optimality of an estimator. One is its accuracy, by which is meant some measure of the concentration of the estimator about the true risk  $R$  [34]. We take as the accuracy of the estimator  $\hat{R}(\alpha)$  its variance  $\text{VAR}\{\hat{R}(\alpha)\} = \frac{V(\alpha)}{N}$ , where  $V(\alpha)$  is the coefficient of variance defined in section 3.2.5 and  $N$  is the sample size. Thus the smaller the coefficient of variance, or the greater the sample size, the greater the accuracy. By the Central Limit Theorem [22], each estimator  $\hat{R}(\alpha)$ ,  $0 \leq \alpha < \alpha_{\max}$  is asymptotically normal with mean  $R$  and variance  $\frac{V(\alpha)}{N}$ . Thus, at least asymptotically, all information about the accuracy of an estimator in the family is contained in its variance.

The other consideration in the optimality of an estimator is the amount of computation it requires. In many problems [39, 1, 23], point evaluations of the conditional densities used in risk estimators are costly. Thus the amount of computation required by an estimator  $\hat{R}(\alpha)$

is taken as the expected number of density evaluations  $NC(\alpha)$  necessary to obtain the estimate, where  $C(\alpha)$  is defined in section 3.2.4 as the expected number of density evaluations per sample and  $N$  is the number of samples.

The estimator in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  with the smallest coefficient of variance  $V(\alpha)$  has the property that it requires the least number of samples to achieve a given accuracy. However, when density evaluations are costly, the size of the sample is not sufficient to characterize the amount of computation required by an estimator  $\hat{R}(\alpha)$ , since the average number of density evaluations  $C(\alpha)$  it requires per sample is also a factor. Thus rather than the optimality criterion of minimum variance we choose the criterion of maximum computational efficiency  $\mathcal{CE}(\alpha)$ . The estimator  $\hat{R}(\alpha^*)$  with maximum computational efficiency has the property that it requires the least amount of computation to achieve a given accuracy [16, 17].

Because of the behavior of the computational efficiency  $\mathcal{CE}(\alpha)$  as a function of  $\alpha$ , maximization may be carried out over a finite number of points. An algorithm to determine  $\alpha^*$  to maximize the computational efficiency  $\mathcal{CE}(\alpha)$  is presented.

The optimal estimator  $\hat{R}(\alpha^*)$  is compared with the existing error count and posterior estimators. It is shown that the more accurate the estimate of the risk, the greater the computational savings will be by using the optimal estimator.

Since in practice it is not possible to maximize the computational efficiency analytically, a technique whereby  $n$  of the total  $N$  samples are used to approximate the optimal estimator is presented. The  $n$  samples should contain enough information on the gross properties of the densi-

ties, such as the closeness of various classes, to closely approximate the optimal estimator. The remaining  $N-n$  samples are used to obtain an accurate estimate of the risk with minimum computation.

#### 4.2 Computational Efficiency: A Criterion for the Optimal Estimator

The computational efficiency  $\mathcal{CE}(\alpha)$  of an estimator  $\hat{R}(\alpha)$  is defined as the inverse of the product of the amount of computation it requires and its accuracy. Since the accuracy of the estimator  $\hat{R}(\alpha)$  is taken as its variance  $\frac{V(\alpha)}{N}$ , and the computational requirements as  $N \times C(\alpha)$ , its average number of density evaluations, we have

##### Definition 4-1

$$\mathcal{CE}(\alpha) = \frac{1}{V(\alpha) \times C(\alpha)} .$$

The optimal estimator in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  is defined as that estimator  $\hat{R}(\alpha^*)$  with maximum computational efficiency. Thus

##### Definition 4-2

The optimal estimator in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  is  $\hat{R}(\alpha^*)$ , where  $\alpha^*$  is such that

$$\max_{0 \leq \alpha < \alpha_{\max}} \mathcal{CE}(\alpha) = \mathcal{CE}(\alpha^*) . \quad (4-1)$$

The optimal estimator  $\hat{R}(\alpha^*)$  has the property that it achieves any given accuracy with a minimum of computation. More precisely, let the sample size  $N_a^*$  be chosen such the estimator  $\hat{R}(\alpha^*)$  based on  $N_a^*$  samples obtains accuracy  $a$ . That is, let  $N_a^*$  be such that

$$\frac{V(\alpha^*)}{N_a^*} = a . \quad (4-2)$$

Let  $\hat{R}(\alpha)$  be any other estimator in the family with sample size  $N$  chosen to obtain the same accuracy  $a$ .

Thus  $N$  is such that

$$\frac{V(\alpha)}{N} = a. \quad (4-3)$$

Then the amount of computation required by  $\hat{R}(\alpha^*)$  based on  $N_a^*$  samples is less than that required by  $\hat{R}(\alpha)$  on  $N$  samples, i.e.

$$N_a^* \times C(\alpha^*) \leq N \times C(\alpha). \quad (4-4)$$

The above is merely the statement that  $\alpha^*$ ,  $N_a^* = \frac{V(\alpha^*)}{a}$  solve the constrained minimization problem (M1)

$$\begin{array}{ll} \text{(M1)} & \text{minimize} \quad N \times C(\alpha) \\ & 0 \leq N \\ & 0 \leq \alpha < \alpha_{\max} \\ & \text{subject to} \quad \frac{V(\alpha)}{N} = a. \end{array}$$

Thus we have

Theorem 4-1

$$\text{Let } \alpha^* \text{ be such that } \mathcal{C}\mathcal{E}(\alpha^*) = \max_{0 \leq \alpha < \alpha_{\max}} \mathcal{C}\mathcal{E}(\alpha)$$

$$\text{and let } N_a^* = \frac{V(\alpha^*)}{a}.$$

Then  $\alpha^*$ ,  $N_a^*$  solve the constrained minimization problem (M1).

Proof:

By definition of  $\alpha^*$

$$\mathcal{C}\mathcal{E}(\alpha^*) \geq \mathcal{C}\mathcal{E}(\alpha) \quad \forall 0 \leq \alpha < \alpha_{\max}. \quad (4-5)$$

By the definition 4-1 of  $\mathcal{C}\mathcal{E}(\alpha)$  we have

$$V(\alpha^*) \times C(\alpha^*) \leq V(\alpha) \times C(\alpha) \quad \forall 0 \leq \alpha < \alpha_{\max}. \quad (4-6)$$

By definition of  $N_a^*$ , we may write  $V(\alpha^*) = a N_a^*$ , thus from (4-6)

$$a N_a^* \times C(\alpha^*) \leq V(\alpha) \times C(\alpha) \quad \forall 0 \leq \alpha < \alpha_{\max}. \quad (4-7)$$

If  $N, \alpha$  are such that  $\frac{V(\alpha)}{N} = a$  then  $V(\alpha) = aN$ . Thus from (4-7)

$$N_a^* \times C(\alpha^*) \leq N \times C(\alpha) \quad \forall \alpha, N \ni \frac{V(\alpha)}{N} = a. \quad (4-8)$$

By symmetry, the optimal estimator  $\hat{R}(\alpha^*)$  also has the property that for a given amount of computation  $b$ ,  $\hat{R}(\alpha^*)$  achieves the greatest accuracy. Thus  $\alpha^*$  and  $N_b^* = \frac{b}{C(\alpha^*)}$  solve the constrained minimization problem (M2).

$$(M2) \quad \begin{array}{ll} \text{minimize} & \frac{V(\alpha)}{N} \\ 0 \leq N & \\ 0 \leq \alpha < \alpha_{\max} & \end{array}$$

subject to  $N \times C(\alpha) = b$ .

#### 4.3 An Algorithm for Maximization of The Computational Efficiency

The optimal estimator  $\hat{R}(\alpha^*)$  from the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  is determined by finding  $\alpha^*$  to maximize the computational efficiency  $CE(\alpha)$ . Because of the behavior of the coefficient of variance  $V(\alpha)$  and the expected number of density evaluations per sample  $C(\alpha)$ , it is only necessary to consider those values of  $\alpha$  that induce changes in the sets  $T_m(\alpha)$  for some  $m=1, 2, \dots, M$ , namely  $\alpha_{t_0}, \alpha_{t_1}, \dots, \alpha_{t_K}$  of definition 3-4, to determine  $\alpha^*$ . This result is proved in the following theorem.

##### Theorem 4-2

$$\max_{0 \leq \alpha < \alpha_{\max}} CE(\alpha) = \max_{i=0, 1, \dots, K} CE(\alpha_{t_i}).$$

##### Proof:

From section 3.2.5,  $V(\alpha)$  is constant  $\forall \alpha \ni \alpha_{t_i} \leq \alpha < \alpha_{t_{i+1}}$ . From section

3.2.4,  $C(\alpha)$  is non-decreasing  $\forall \alpha \ni \alpha_{t_i} \leq \alpha < \alpha_{t_{i+1}}$ . Thus

$CE(\alpha) = \frac{1}{V(\alpha) \times C(\alpha)}$  is non-increasing  $\forall \alpha \ni \alpha_{t_i} \leq \alpha < \alpha_{t_{i+1}}$

Therefore,

$$\max_{\alpha_{t_i} \leq \alpha < \alpha_{t_{i+1}}} CE(\alpha) = CE(\alpha_{t_i}) \quad (4-9)$$

Finally

$$\begin{aligned} \max_{0 \leq \alpha < \alpha_{\max}} CE(\alpha) &= \max_{i=0,1,\dots,K} \max_{\alpha_{t_i} \leq \alpha < \alpha_{t_{i+1}}} CE(\alpha) \\ &= \max_{i=0,1,\dots,K} CE(\alpha_{t_i}). \end{aligned} \quad (4-10)$$

Since it has been observed that the coefficient of variance  $V(\alpha)$  is non-decreasing in  $\alpha$ , we state the following corollary. In this case,  $CE(\alpha)$  may be maximized over  $\alpha_{q_0}, \alpha_{q_1}, \dots, \alpha_{q_J}$  of definition 3-7, the subset of  $\alpha_{t_0}, \alpha_{t_1}, \dots, \alpha_{t_K}$  which induce changes in the sets  $Q_m(\alpha)$  for some  $m=1,2,\dots,M$ . The convenience is that in general  $J < K$ , thus maximization may be carried out over fewer points.

#### Corollary 4-2

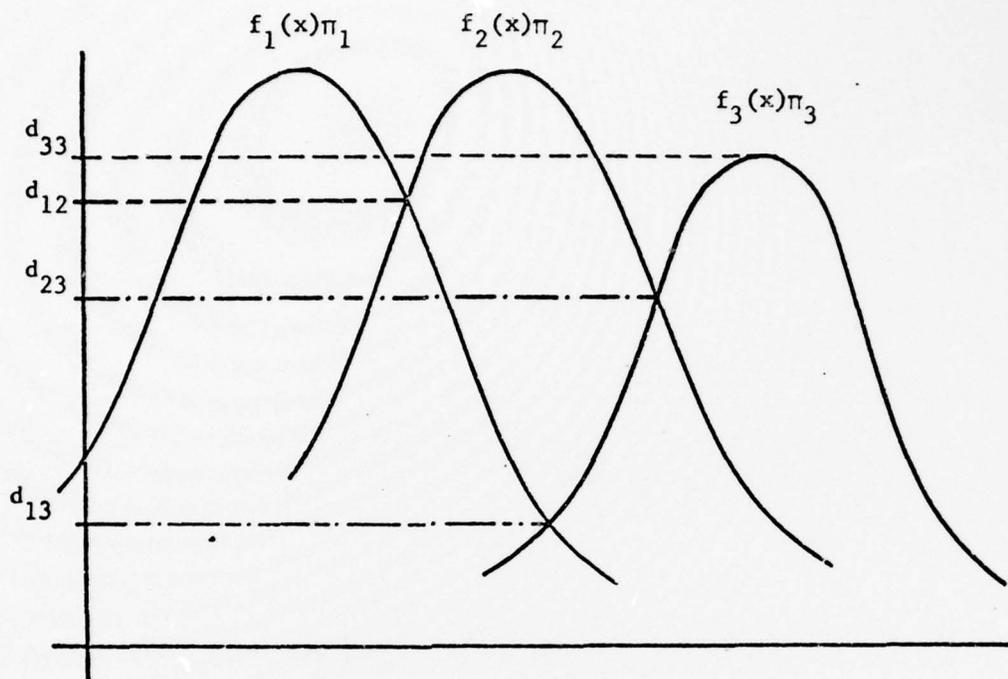
If  $V(\alpha)$  is non-decreasing in  $\alpha$  then

$$\max_{0 \leq \alpha < \alpha_{\max}} CE(\alpha) = \max_{i=0,1,\dots,J} CE(\alpha_{q_i})$$

Proof:

Follows since  $C(\alpha)$  and  $V(\alpha)$  are non-decreasing on  $\alpha_{q_i} \leq \alpha < \alpha_{q_{i+1}}$ .

Thus the problem of maximizing the computational efficiency  $CE(\alpha)$  is reduced to finding the points  $\alpha_{t_0}, \alpha_{t_1}, \dots, \alpha_{t_K}$  and evaluating  $CE(\alpha)$  at these points. We now describe a convenient method to find  $\alpha_{t_0}, \alpha_{t_1}, \dots, \alpha_{t_K}$  and



$d_{ij}$	$\alpha_{t_i}$	$T_1$	$T_2$	$T_3$
-	$\alpha_{t_0}$	123	123	123
$d_{13}$	$\alpha_{t_1}$	12	123	23
$d_{23}$	$\alpha_{t_2}$	12	12	3
$d_{12}$	$\alpha_{t_3}$	1	2	3
$d_{33}$	$\alpha_{\max}$	-	-	-

Figure 4-1. The points  $d_{rs}$ ,  $r=1,2,3, s=r...3$  which split classes  $r$  and  $s$ , their association with  $\alpha_{t_0} \dots \alpha_{t_3}$  and the sets  $T_m$ .

the resulting sets  $T_m(\alpha_{t_i})$ ,  $m=1,2,\dots,M$ ,  $i=0,1,\dots,K$ .

Let  $d_{rs}$  be the value of  $\alpha$  which splits classes  $r$  and  $s$ , defined as follows.

Definition 4-3

$$d_{rs} = \max_{x \in S} \min \{f_r(x)\pi_r, f_s(x)\pi_s\}$$

$$r=1,2,\dots,M, s=r, r+1,\dots,M.$$

Note that for  $\alpha < d_{rs}$ ,  $r \in T_s(\alpha)$  and  $s \in T_r(\alpha)$ . This follows since if  $\alpha < \max_{x \in S} \min \{f_r(x)\pi_r, f_s(x)\pi_s\}$  then  $\exists x \in S$  such that  $f_r(x)\pi_r > \alpha$  and  $f_s(x)\pi_s > \alpha$ . For  $\alpha \geq d_{rs}$ ,  $r \notin T_s(\alpha)$  and  $s \notin T_r(\alpha)$ . Thus  $d_{rs}$  is the smallest value of  $\alpha$  that splits classes  $r$  and  $s$ , in the sense that  $\forall \alpha \geq d_{rs}$ ,  $r \notin T_s(\alpha)$  and  $s \notin T_r(\alpha)$ . Figure 4-1 shows three joint densities and the values of  $d_{12}$ ,  $d_{13}$  and  $d_{23}$ .

For simplification, assume that the points  $d_{rs}$ ,  $r = 1,2,\dots,M-1$ ,  $s = r+1,\dots,M$  are distinct. Then the values  $\alpha_{t_1}, \alpha_{t_2}, \dots, \alpha_{t_K}$  and  $\alpha_{\max}$  may be obtained from  $d_{rs}$ ,  $r = 1,2,\dots,M$ ,  $s = r,r+1,\dots,M$  as follows.

Definition 4-4

Order the values of  $d_{rs}$  in increasing order as follows

$$d_{r_0 s_0} = 0$$

Do  $i=0$  by 1

$$d_{r_{i+1} s_{i+1}} = \min_{\substack{r,s \\ \exists d_{rs} > d_{r_i s_i}}} d_{rs}$$

Stop if  $r_{i+1} = s_{i+1}$

End.

Theorem 4-3

$$\alpha_{t_i} = d_{r_i, s_i} \quad i=0, 1, \dots, K$$

$$\alpha_{\max} = d_{r_{K+1}, s_{K+1}}$$

Proof:

By induction.

$$\alpha_{t_0} = d_{r_0, s_0} = 0 \text{ by definition.}$$

$$\text{Suppose } \alpha_{t_i} = d_{r_i, s_i}, \quad i=1, 2, \dots, k < K.$$

By definition,  $d_{r_{k+1}, s_{k+1}}$  is the smallest value of  $\alpha > d_{r_k, s_k}$  such that  $s_{k+1} \notin T_r(\alpha)$  and  $r_{k+1} \notin T_{s_{k+1}}(\alpha)$ . Since  $d_{r_k, s_k} = \alpha_{t_k}$  by the induction hypothesis,  $d_{r_{k+1}, s_{k+1}}$  is the smallest value of  $\alpha > \alpha_{t_k}$  such at  $T_m(\alpha) \neq T_m(\alpha_{t_k})$  for some  $m$  (namely  $m=r_{k+1}, s_{k+1}$ ). Thus  $d_{r_{k+1}, s_{k+1}} = \alpha_{t_{k+1}}$ .

$$\begin{aligned} \text{Also, by definitions 3-3 and 4-3, } \alpha_{\max} &= \min_{1 \leq l \leq M} \max_{x \in S} f_l(x) \prod_l \\ &= \min_{1 \leq l \leq M} d_{r_l, s_l} = d_{r_{K+1}, s_{K+1}}. \end{aligned}$$

The sets  $T_m(\alpha_{t_i})$   $m=1, 2, \dots, M$ ,  $i=0, 1, \dots, K$  may also be determined from the ordered values  $d_{rs}$  by the following corollary.

Corollary 4-3

$$T_m(\alpha_{t_0}) = \{i \mid \exists x \ni f_i(x)\pi_i > 0 \text{ and } f_m(x)\pi_m > 0\}$$

Do  $i=0$  to  $K-1$ .

$$T_m(\alpha_{t_{i+1}}) = T_m(\alpha_{t_i}) \forall m \neq r_i, m \neq s_i$$

$$T_{r_i}(\alpha_{t_{i+1}}) = T_{r_i}(\alpha_{t_i}) - s_i \text{ (delete } s_i \text{ from } T_{r_i}(\alpha_{t_i}))$$

$$T_{s_i}(\alpha_{t_{i+1}}) = T_{s_i}(\alpha_{t_i}) - r_i \text{ (delete } r_i \text{ from } T_{s_i}(\alpha_{t_i})).$$

Figure 4-1 shows the values  $d_{12}$ ,  $d_{13}$ ,  $d_{23}$  and  $d_{33}$  and their correspondence to  $\alpha_{t_0} \dots \alpha_{t_K}$ ,  $\alpha_{\max}$  and the sets  $T_m(\alpha_{t_i})$ ,  $m=1,2\dots M$ ,  $i=0,1\dots K$ .

Assuming that the values  $d_{rs}$ ,  $r=1,2\dots M$ ,  $s=r, r+1\dots M$  have been computed from definition 4-3 and placed in increasing order in correspondence with  $\alpha_{t_0}, \alpha_{t_1} \dots \alpha_{t_K}$  as in definition 4-4 and theorem 4-3, an algorithm to determine  $\alpha^*$  to maximize the computational efficiency  $CE(\alpha)$  is as follows.

Algorithm A

Let  $\alpha_{t_0} = 0$ .

1) For  $m=1,2\dots M$  let

$$T_m(\alpha_{t_0}) = \{l \mid \exists x \ni f_m(x)\pi_m > \alpha_{t_0}, f_l(x)\pi_l > \alpha_{t_0}\}$$

2) For  $m=1,2\dots M$  let

$$Q_m(\alpha_{t_0}) = \bigcup_{q \in T_m(\alpha_{t_0})} T_q(\alpha_{t_0}).$$

3) Compute  $CE(\alpha_{t_0}) = \frac{1}{C(\alpha_{t_0}) \times V(\alpha_{t_0})}$ .

4) Max =  $CE(\alpha_{t_0})$ ,  $\alpha^* = \alpha_{t_0}$ .

Do for  $i=1$  to  $K$ .

5) For  $m=1, 2, \dots, M$  let

$$T_m(\alpha_{t_i}) = T_m(\alpha_{t_{i-1}}) \quad \forall m \neq r_i, s_i$$

$$T_{r_i}(\alpha_{t_i}) = T_{r_i}(\alpha_{t_{i-1}}) - s_i$$

$$T_{s_i}(\alpha_{t_i}) = T_{s_i}(\alpha_{t_{i-1}}) - r_i$$

6) For  $m=1, 2, \dots, M$  let

$$Q_m(\alpha_{t_i}) = \bigcup_{q \in T_m(\alpha_{t_i})} T_q(\alpha_{t_i})$$

7) Compute  $CE = \frac{1}{C(\alpha_{t_i}) \times V(\alpha_{t_i})}$ .

8) If  $CE(\alpha_{t_i}) > \text{Max}$  then

$$\text{Max} = CE(\alpha_{t_i}), \quad \alpha^* = \alpha_{t_i}$$

Note that if  $V(\alpha)$  is non-decreasing in  $\alpha$ , by corollary 4-2, steps 7) and 8) need only be done for those  $i$  such that  $\exists m$  such that  $Q_m(\alpha_{t_i}) \neq Q_m(\alpha_{t_{i-1}})$ . Thus  $CE(\alpha)$  need only be evaluated at those values  $\alpha_{q_0}, \alpha_{q_1}, \dots, \alpha_{q_J}$  which induce changes in the sets  $Q_m(\alpha)$ ,  $m=1, 2, \dots, M$ .

#### 4.4 Comparison of the Optimal Estimator With the Error Count and Posterior Estimators

We first compare the optimal estimator  $\hat{R}(\alpha^*)$  to the posterior estimator  $\hat{R}(p)$ , defined in section 3.2.1, on the basis of relative computational efficiency. The posterior estimator requires for each sample  $X_j$ ,  $j=1, 2, \dots, N$ , the evaluation of all  $M$  conditional densities  $f_{\ell}(X_j)$ ,  $\ell=1, 2, \dots, M$ , a total of  $N \times M$  density evaluations. The variance of the

posterior estimator is, from (3-9)

$$\text{VAR}\{\hat{R}(p)\} = \frac{\text{VAR}\{r(X)\}}{N} . \quad (4-11)$$

Let the coefficient of variance  $V(p)$  be

$$V(p) = \text{VAR}\{r(X)\} . \quad (4-12)$$

Then the computational efficiency  $\mathcal{CE}(p)$  of the posterior estimator is

$$\mathcal{CE}(p) = \frac{1}{M \times V(p)} . \quad (4-13)$$

The computational efficiency of the optimal estimator relative to the posterior estimator,  $\mathcal{RCE}(\alpha^*, p)$  is defined by

$$\mathcal{RCE}(\alpha^*, p) = \frac{\mathcal{CE}(\alpha^*)}{\mathcal{CE}(p)} = \frac{M \times V(p)}{C(\alpha^*) \times V(\alpha^*)} . \quad (4-14)$$

The following theorem states that the computational efficiency of the optimal estimator is greater than or equal to that of the posterior estimator.

Theorem 4-4

$$\mathcal{RCE}(\alpha^*, p) \geq 1$$

Proof:

The estimator  $\hat{R}(0)$  in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  is equivalent to  $\hat{R}(p)$  in the sense that

$$\hat{R}(0) = \hat{R}(p) . \quad (4-15)$$

Thus  $\hat{R}(0)$  and  $\hat{R}(p)$  have the same coefficient of variance

$$V(0) = V(p) . \quad (4-16)$$

However, it may be the case (if the conditional densities have finite

support) that  $\hat{R}(0)$  may be computed with fewer than  $M$  conditional density evaluations per sample. Thus

$$C(0) \leq M. \quad (4-17)$$

From (4-16) and (4-17), we have

$$V(0) \times C(0) \leq V(p) \times C(p) \quad (4-18)$$

and by definition of computational efficiency,

$$CE(0) \geq CE(p). \quad (4-19)$$

By definition 4-2 of  $\alpha^*$

$$CE(\alpha^*) \geq CE(\alpha) \quad \forall 0 \leq \alpha < \alpha_{\max}. \quad (4-20)$$

Thus

$$CE(\alpha^*) \geq CE(p), \quad (4-21)$$

and finally

$$RCE(\alpha^*, p) = \frac{CE(\alpha^*)}{CE(p)} \geq 1. \quad (4-22)$$

The computational efficiency of the optimal estimator relative to the posterior,  $RCE(\alpha^*, p)$  has the interpretation that if the sample size  $N_p$  for the posterior estimator and  $N_*$  for the optimal estimator are chosen so that both estimators have the same accuracy (variance), then the posterior estimator will require  $RCE(\alpha^*, p)$  times the number of density evaluations required by the optimal estimator.

To see this, let the sample sizes  $N_*$  and  $N_p$  be chosen so that

$$\frac{V(\alpha^*)}{N_*} = \frac{V(p)}{N_p}. \quad (4-23)$$

The amount of computation, expressed as the average number of density evaluations, required by the optimal estimator is  $N_* \times C(\alpha^*)$ . The number of density evaluations required by the posterior to achieve the same accuracy is  $N_p \times M$ . From (4-23), we have that

$$\begin{aligned} N_p \times M &= \frac{V(p) \times N_*}{V(\alpha^*)} \times M = \frac{V(p) \times M}{V(\alpha^*) \times C(\alpha^*)} (N_* \times C(\alpha^*)) \\ &= \frac{CE(\alpha^*)}{CE(p)} (N_* \times C(\alpha^*)) = RCE(\alpha^*, p) (N_* \times C(\alpha^*)) . \end{aligned} \quad (4-24)$$

The posterior estimator requires  $RCE(\alpha^*, p)$  times the number of density evaluations required by the optimal estimator to obtain the same accuracy. Thus by using the optimal estimator  $\hat{R}(\alpha^*)$ , we have saved ourselves, on the average,  $S(\alpha^*, p)$  density evaluations, where

$$S(\alpha^*, p) = (RCE(\alpha^*, p) - 1) C(\alpha^*) \times N_* . \quad (4-25)$$

From (4-25), it is clear that the more accurate an estimate of the risk desired, that is, the larger  $N_*$ , the greater the savings in computation  $S(\alpha^*, p)$ .

The optimal estimator  $\hat{R}(\alpha^*)$  compares even more favorably to the error count estimator  $\hat{R}(ec)$ , defined in section 3.2.1. The variance of the error count estimator is, from (3-3),

$$VAR\{\hat{R}(ec)\} = \frac{R(1-R)}{N} . \quad (4-26)$$

Since the error count estimator requires evaluation of the conditional density  $f_l(X_j)$  for each sample  $X_j$ ,  $j=1, 2, \dots, N$  and for each class  $l=1, 2, \dots, M$ , the computational efficiency of the error count estimator  $CE(ec)$  is given

by

$$CE(ec) = \frac{1}{M \times R(1-R)} \quad (4-27)$$

The computational efficiency of the error count estimator is less than that of the posterior. From (3-11), we have that

$$V(p) \leq R(1-R) - \frac{R}{M} < R(1-R). \quad (4-28)$$

Thus

$$CE(ec) = \frac{1}{M \times R(1-R)} < \frac{1}{M \times V(p)} = CE(p) \quad (4-29)$$

The computational efficiency of the optimal estimator relative to the error count estimator,  $RCE(\alpha^*, ec)$ , is given by

$$RCE(\alpha^*, ec) = \frac{CE(\alpha^*)}{CE(ec)} = \frac{M \times R(1-R)}{C(\alpha^*) \times V(\alpha^*)} \quad (4-30)$$

From (4-29) and theorem 4-4 we have

$$RCE(\alpha^*, ec) > RCE(\alpha^*, p) \geq 1 \quad (4-31)$$

If the error count estimator  $\hat{R}(ec)$  is a member of the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  and if  $V(\alpha)$  is non-decreasing in  $\alpha$ , we have

$$V(\alpha^*) \leq R(1-R) \quad (4-32)$$

In this case, a lower bound on the computational efficiency of the optimal estimator relative to the posterior estimator is

$$RCE(\alpha^*, e) \geq \frac{M \times R(1-R)}{C(\alpha^*) \times R(1-R)} = \frac{M}{C(\alpha^*)} \quad (4-33)$$

Thus if (4-32) holds, the error count estimator requires at least  $M/C(\alpha^*)$  times the number of conditional density evaluations required by the opti-

mal estimator to obtain the same accuracy.

The number of density evaluations saved by using the optimal estimator rather than the error count,  $\mathfrak{S}(\alpha^*, ec)$  is

$$\begin{aligned} \mathfrak{S}(\alpha^*, ec) &= (\mathcal{RCE}(\alpha^*, ec) - 1) C(\alpha^*) \times N_* \\ &\geq \left( \frac{M}{C(\alpha^*)} - 1 \right) C(\alpha^*) \times N_* \\ &= (M - C(\alpha^*)) N_* \end{aligned} \quad (4-34)$$

Thus the more accurate an estimate of the risk desired (the larger  $N_*$ ) the greater savings. Also, the smaller  $C(\alpha^*)$  relative to the total number  $M$  of classes, the greater the savings. One would expect  $M \gg C(\alpha^*)$  for a large number  $M$  of classes which tend to form several small clusters.

#### 4.5 Approximation of the Optimal Estimator

The optimal estimator  $\hat{R}(\alpha^*)$  in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  is determined by finding  $\alpha^*$  to maximize the computational efficiency  $\mathcal{CE}(\alpha)$ . However, if we had enough information to maximize the computational efficiency analytically, we could evaluate the Bayes risk  $R$  analytically. We propose that a subset of the data, say  $\{(X_1, \theta_1), (X_2, \theta_2), \dots, (X_n, \theta_n)\}$ , where  $n \ll N$ , be used to approximate the optimal estimator. The remaining  $N-n$  samples are used in the approximated optimal estimator to obtain an accurate estimate of the Bayes risk efficiently.

Recall algorithm A in section 4.3 for finding  $\alpha^*$  to maximize the computational efficiency  $\mathcal{CE}(\alpha)$ . In order to use this algorithm, we need to know the points  $d_{rs}$ ,  $r=1, 2, \dots, M$ ,  $s=r, r+1, \dots, M$ . and the value of the

computational efficiency at these points. Since in practice these values are not known, we propose they be approximated on the basis of the  $n$  samples  $\{(X_1, \theta_1) \dots (X_n, \theta_n)\}$  as follows.

An approximation to the computational efficiency  $\mathcal{CE}(\alpha)$  at any given  $\alpha$  is formed as

$$\hat{\mathcal{C}}\mathcal{E}(\alpha) = \frac{1}{\hat{V}(\alpha) \times \hat{C}(\alpha)}, \quad (4-35)$$

where  $\hat{V}(\alpha)$  and  $\hat{C}(\alpha)$  are unbiased estimates of  $V(\alpha)$  and  $C(\alpha)$  given by

$$\hat{C}(\alpha) = \sum_{\ell=1}^M |Q_{\ell}(\alpha)| (\pi_{\ell} - \hat{U}_{\ell}(\alpha)) + M \hat{U}_{\ell}(\alpha) \quad (4-36)$$

where

$$\hat{U}_{\ell}(\alpha) = \frac{1}{n} \sum_{j=1}^n (1 - I_{\ell}(X_j, \alpha)) \frac{f_{\ell}(X_j) \pi_{\ell}}{f(X_j)} \quad (4-37)$$

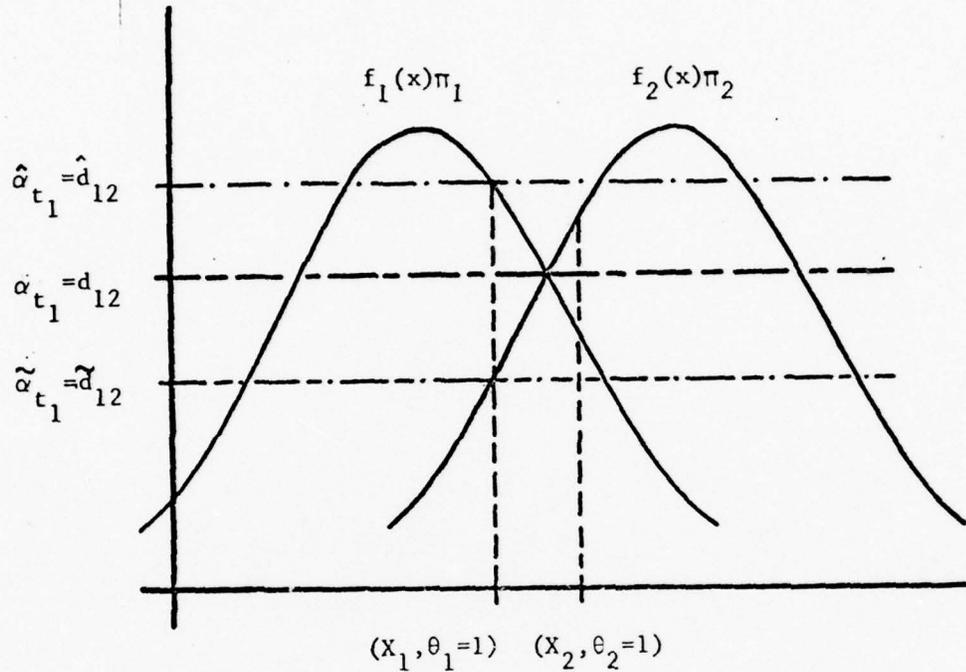
and

$$\begin{aligned} \hat{V}(\alpha) = & \frac{1}{n-1} \sum_{j=1}^n \left[ \sum_{m \in T_{\theta_j}(\alpha)} \mathcal{E}_m(X_j) \frac{f_m(X_j) \pi_m}{\sum_{\ell \in T_m(\alpha)} f_{\ell}(X_j) \pi_{\ell}} \right. \\ & \left. - \left( \frac{1}{n} \sum_{k=1}^n \sum_{m \in T_{\theta_k}(\alpha)} \mathcal{E}_m(X_k) \frac{f_m(X_k) \pi_m}{\sum_{\ell \in T_m(\alpha)} f_{\ell}(X_k) \pi_{\ell}} \right)^2 \right] \end{aligned} \quad (4-38)$$

The points  $d_{rs}$ ,  $r=1, 2, \dots, M$ ,  $s=r, r+1, \dots, M$  might be approximated by

$$\tilde{d}_{rs} = \max_{1 \leq j \leq n} \min \{f_r(X_j) \pi_r, f_s(X_j) \pi_s\}. \quad (4-39)$$

Once the values  $\tilde{d}_{rs}$ ,  $r=1, 2, \dots, M$ ,  $s=r, r+1, \dots, M$  have been ordered and put into correspondence with the points  $\tilde{\alpha}_{t_0}, \tilde{\alpha}_{t_1}, \dots, \tilde{\alpha}_{t_K}$  as in theorem 4-3,



$\alpha$	$T$	$\hat{T}$	$Q$	$\hat{Q}$
$\tilde{\alpha}_{t_1}$	$T_1(\tilde{\alpha}_{t_1}) = \{1, 2\}$	$\hat{T}_1(\tilde{\alpha}_{t_1}) = \{1\}$	$Q_1(\tilde{\alpha}_{t_1}) = \{1, 2\}$	$\hat{Q}_1(\tilde{\alpha}_{t_1}) = \{1\}$
	$T_2(\tilde{\alpha}_{t_1}) = \{1, 2\}$	$\hat{T}_2(\tilde{\alpha}_{t_1}) = \{2\}$	$Q_2(\tilde{\alpha}_{t_1}) = \{1, 2\}$	$\hat{Q}_2(\tilde{\alpha}_{t_1}) = \{2\}$

$\alpha$	$T$	$\hat{T}$	$Q$	$\hat{Q}$
$\hat{\alpha}_{t_1}$	$T_1(\hat{\alpha}_{t_1}) = \{1\}$	$\hat{T}_1(\hat{\alpha}_{t_1}) = \{1\}$	$Q_1(\hat{\alpha}_{t_1}) = \{1\}$	$\hat{Q}_2(\hat{\alpha}_{t_1}) = \{1\}$
	$T_2(\hat{\alpha}_{t_1}) = \{2\}$	$\hat{T}_2(\hat{\alpha}_{t_1}) = \{2\}$	$Q_2(\hat{\alpha}_{t_1}) = \{2\}$	$\hat{Q}_q(\hat{\alpha}_{t_1}) = \{2\}$

Figure 4-2. Estimates  $\hat{d}_{12}$  and  $\tilde{d}_{12}$  of  $d_{12}$  based on one sample  $(x_1, \theta_1=1)$ .  $\tilde{d}_{12}$  underestimates  $d_{12}$  which results in sets  $\hat{T}_m$  that are smaller than the true sets  $T_m$ . The overestimate  $\hat{d}_{12}$  solves this problem.

algorithm A may be performed by substituting the approximated values for the true values. As a result, a value  $\tilde{\alpha}^*$  which maximizes the approximate computational efficiency  $\hat{C}\mathcal{E}$  is obtained.

However, the following difficulty arises. Algorithm A determines the sets  $\hat{T}_m(\tilde{\alpha}_{t_i})$ ,  $m=1,2,\dots,M$ ,  $i=0,1,\dots,K$  as in corollary 4-3. But since

$$\tilde{d}_{rs} \leq d_{rs} \quad r=1,2,\dots,M, s=r, r+1,\dots,M \quad (4-40)$$

the sets  $\hat{T}_m(\tilde{\alpha}_{t_i})$ ,  $m=1,2,\dots,M$ ,  $i=0,1,\dots,K$  which result from corollary 4-3 using the approximated values  $\tilde{d}_{rs}$  and  $\tilde{\alpha}_{t_i}$  have the property that

$$\hat{T}_m(\tilde{\alpha}_{t_i}) \subset T_m(\tilde{\alpha}_{t_i}). \quad (4-41)$$

Thus the approximated set  $\hat{T}_m(\tilde{\alpha}_{t_i})$  of classes " $\tilde{\alpha}_{t_i}$ -close" to class  $m$  is smaller than the true set  $T_m(\tilde{\alpha}_{t_i})$  of classes " $\tilde{\alpha}_{t_i}$ -close" to class  $m$ .

Figure 4-2 shows this behavior for 2 classes, with  $\tilde{d}_{12} = \tilde{\alpha}_{t_1}$  approximated with  $n=1$  sample  $X_1, e_1 = 1$ .

The result of this is that if the estimator  $c\hat{R}(\tilde{\alpha})$ , in computational form, is used with the approximated sets  $\hat{T}_m(\tilde{\alpha})$ ,  $m=1,2,\dots,M$ , a biased estimate of the risk results. The reason is that in its computational form,  $c\hat{R}(\tilde{\alpha})$  uses the modified error function  $\mathcal{E}_m(X, \hat{Q}_\theta(\tilde{\alpha}))$  for  $m \in \hat{T}_\theta(\tilde{\alpha})$  whenever  $f_\theta(X)\pi_\theta > \tilde{\alpha}$ . Although by theorem 3-2 it is true that

$$\forall m \in T_\theta(\alpha), \mathcal{E}_m(X) = \mathcal{E}_m(X, Q_\theta(\alpha))$$

(4-42)

whenever  $f_\theta(X)\pi_\theta > \tilde{\alpha}$

it is not true in general that

$$\forall m \in \hat{T}_\theta(\tilde{\alpha}), \mathcal{E}_m(X) = \mathcal{E}_m(X, \hat{Q}_\theta(\tilde{\alpha}))$$

(4-43)

whenever  $f_\theta(X)\pi_\theta > \tilde{\alpha}$

since  $\hat{Q}_\theta(\tilde{\alpha})$  may be smaller than  $Q_\theta(\tilde{\alpha})$ .

#### Example 4-1

In figure 4-2, the values  $\tilde{\alpha}_{t_1}$  and  $\hat{T}_m(\tilde{\alpha}_{t_1})$ ,  $m=1,2$  were approximated with the sample  $X_1, \theta_1 = 1$ . If sample  $X_2, \theta_2 = 1$  were used in the computational form  $c\hat{R}(\tilde{\alpha}_{t_1})$ , since  $f_1(X_2)\pi_1 > \tilde{\alpha}_{t_1}$  the modified error function would be used. But  $\mathcal{E}_1(X_2, \hat{Q}_1(\tilde{\alpha}_{t_1})) = 0$  while  $\mathcal{E}_1(X_2) = 1$ .

Examples indicate that the sets  $\hat{T}_m(\tilde{\alpha}_{t_i})$ ,  $m=1,2\dots M$ ,  $i=0,1\dots K$  approximate well the true sets  $T_m(\alpha_{t_i})$ ,  $m=1,2\dots M$ ,  $i=0,1\dots K$  (see section 4.6) The ad-hoc method employed here is to overestimate the values

$d_{r_i s_i}$ ,  $i=0,1\dots K+1$  by

$$\hat{d}_{r_i s_i} = f_{r_i}(X)\pi_{r_i}$$

(4-44)

whenever  $\tilde{d}_{r_i s_i} = f_{s_i}(X)\pi_{s_i}$ .

The value  $\hat{d}_{12}$  is illustrated in figure 4-2. Note that in this case,

$$T_m(\hat{\alpha}_{t_1}) = \hat{T}_m(\hat{\alpha}_{t_1}) \quad m=1,2. \quad (4-45)$$

Let  $\hat{\alpha}^*$  be the approximated optimal determined by Algorithm A, with the approximated values  $\hat{C}\mathcal{E}$  and  $\hat{\alpha}_{t_i}$  (determined from  $\hat{d}_{r_i s_i}$ ) substituted for the true values. Thus  $\hat{\alpha}^*$  and sets  $\hat{T}_m(\hat{\alpha}^*)$ ,  $m=1,2,\dots,M$  determine the approximate optimal estimator  $\hat{R}(\hat{\alpha}^*)$ .

In determining the optimal  $\hat{\alpha}^*$ , all conditional densities

$$f_{\ell}(X_j) \quad \ell=1,2,\dots,M, \quad j=1,2,\dots,n \quad (4-46)$$

have been evaluated. Since the posterior estimator  $\hat{R}(p)$  based on  $n$  samples is a by-product of Algorithm A, and since the posterior estimator has the smallest variance (theorem 3-1) we may as well incorporate it into the approximated optimal estimator.

Thus, let the final estimator  $\hat{R}(f)$  be given by

$$\hat{R}(f) = \frac{n}{N} \hat{R}(p) + \frac{(N-n)}{N} \hat{R}(\hat{\alpha}^*) \quad (4-47)$$

where  $\hat{R}(p)$  is the posterior estimator based on the  $n$  samples  $\{(X_1, \theta_1) \dots (X_n, \theta_n)\}$  used to determine  $\hat{\alpha}^*$  and  $\hat{R}(\hat{\alpha}^*)$  is the approximated optimal estimator based on the remaining  $N-n$  samples  $\{(X_{n+1}, \theta_{n+1}), \dots, (X_N, \theta_N)\}$ .

The final estimator  $\hat{R}(f)$  requires  $M$  density evaluations for each of the first  $n$  samples and an average of  $C(\hat{\alpha}^*)$  for the remaining  $N-n$  samples. Its variance is given by

$$\text{VAR}\{\hat{R}(f)\} = \frac{n}{N^2} V(p) + \frac{(N-n)}{N^2} V(\hat{\alpha}^*) \quad (4-48)$$

Thus the computational efficiency  $C\mathcal{E}(f)$  of the estimator is

$$C\mathcal{E}(f) = \frac{1}{(nM + (N-n)C(\hat{\alpha}^*)) \left( \frac{n}{N^2} V(p) + \frac{(N-n)}{N^2} V(\hat{\alpha}^*) \right)} \quad (4-49)$$

Assume that the  $n$  samples contain enough information so that the approximated optimal estimator  $\hat{R}(\hat{\alpha}^*)$  is close to the true optimal estimator  $\hat{R}(\alpha^*)$ . That is, assume

$$CE(\alpha^*) = \frac{1}{V(\alpha^*) \times C(\alpha^*)} \approx \frac{1}{V(\hat{\alpha}^*) \times C(\hat{\alpha}^*)} = CE(\hat{\alpha}^*). \quad (4-50)$$

Also, let us assume the approximation is close enough so that

$$CE(\hat{\alpha}^*) \geq CE(p). \quad (4-51)$$

Then from (4-49) and (4-51),

$$CE(\hat{\alpha}^*) \geq CE(f) \geq CE(p), \quad (4-52)$$

the lower bound  $CE(p)$  for  $CE(f)$  being obtained when  $n=N$  and the upper bound  $CE(\hat{\alpha}^*)$  when  $n=0$ . The best case would be when  $\alpha^*$  were known a priori and  $n=0$ .

The computational efficiency of the final estimator relative to the posterior estimator,  $\mathcal{R}CE(f,p)$  is

$$\mathcal{R}CE(f,p) = \frac{CE(f)}{CE(p)} = \frac{MV(p)}{(nM + (N-n)C(\hat{\alpha}^*)) \left( \frac{n}{N^2} V(p) + \frac{(N-n)}{N^2} V(\hat{\alpha}^*) \right)} \quad (4-53)$$

which is greater than one provided (4-51) holds.

Let us now compare the final estimator to the posterior estimator in terms of the number of density evaluations saved. As in section 4.4 let  $\mathcal{S}(f,p)$  be the number of density evaluations one would save by using the final estimator on  $N$  samples rather than the posterior based on  $N_p$  samples, where  $N_p$  is chosen so that both have same accuracy (variance).

Then

$$\mathcal{S}(f,p) = (\mathcal{R}CE(f,p) - 1) (nM + (N-n)C(\hat{\alpha}^*)). \quad (4-54)$$

From theorem 3-1, we have

$$V(p) \leq V(\hat{\alpha}^*). \quad (4-55)$$

Thus from (4-53), (4-54) and (4-55),

$$\mathfrak{S}(f,p) \geq N \left( \frac{MxV(p) - V(\hat{\alpha}^*)C(\hat{\alpha}^*)}{V(\hat{\alpha}^*)} \right) - n(M-C(\hat{\alpha}^*)). \quad (4-56)$$

From (4-56), the greatest savings result when  $N$  is large (an accurate risk estimate is desired) and  $n$  is small. However, it is important that  $n$  be large enough to closely approximate the optimal estimator to assure that (4-51) holds.

It can be shown that if

$$V(\hat{\alpha}^*) \leq R(1-R) \quad (4-57)$$

then the average savings in number of density evaluations by using the final estimator rather than the error count estimator,  $\mathfrak{S}(f,ec)$  for an estimate of the same accuracy, is bounded below by

$$\mathfrak{S}(f,ec) \geq (N-n)(M-C(\hat{\alpha}^*)). \quad (4-58)$$

Of course, it takes a certain amount of work, over and above the  $Mn$  density evaluations, to approximate the optimal estimator using the first  $n$  samples. The average number of density evaluations saved by using the final estimator rather than one of the existing estimators must be compared to this overhead. If the density evaluations are costly and the number saved is large, a net savings in work should be realized.

#### 4.6 Examples

Consider the five Gaussian classes of example 1 described on page A-1 of the appendix. Page A-2 gives for these densities the values  $d_{r_i s_i}$ ,

$i=0,1,\dots,10$ , their correspondence with  $\alpha_{t_i}$ ,  $i=0,1,\dots,10$  and  $\alpha_{q_i}$ ,  $i=0,1,\dots,5$ , and the resulting sets  $T_m(\alpha_{t_i})$ ,  $Q_m(\alpha_{t_i})$ ,  $m=1,2,\dots,5$ ,  $i=0,1,\dots,10$ . The computational efficiency  $CE(\alpha_{t_i})$ ,  $i=0,1,\dots,10$  is given on page A-3. The computational efficiency is maximized for  $\alpha^* = \alpha_{t_6}$  and  $CE(\alpha^*) = 20.99$ . From page A-2, we have  $T_1(\alpha^*) = \{1,2\} = T_2(\alpha^*)$ ,  $T_3(\alpha^*) = \{3,4,5\} = T_4(\alpha^*) = T_5(\alpha^*)$ , and  $Q_i(\alpha^*) = T_i(\alpha^*)$   $i=1,2,\dots,5$ . These sets represent the natural clustering of the classes.

On page A-6, the computational efficiency of the optimal estimator relative to the posterior is given by  $RCCE(\alpha^*, p) = 1.92$ . Thus to achieve the same accuracy, the posterior estimator would require on the average 1.92 times the number of density evaluations required by the optimal estimator. The computational efficiency of the optimal estimator relative to the error count estimator is  $RCCE(\alpha^*, ec) = 24.07$ , thus the error count would require 24.07 times the computation of the optimal for the same accuracy.

Also on page A-6 are the average number of density evaluations saved by using the optimal estimator, rather than the error count or posterior, for various sample sizes  $N_*$  for the optimal estimator. For example, if the optimal estimator is formed using  $N_*=400$  samples, in order to achieve the same accuracy the posterior would require on the average 956 more density evaluations and the error count 23,992 more.

Page A-7 gives the approximated values  $\hat{d}_{r_i s_i}$ ,  $i=0,1,\dots,10$ , their correspondence with  $\hat{\alpha}_{t_i}$ ,  $i=0,1,\dots,10$  and  $\hat{\alpha}_{q_i}$ ,  $i=0,1,\dots,5$  and the resulting sets  $T_m(\hat{\alpha}_{t_i})$ ,  $\hat{Q}_m(\hat{\alpha}_{t_i})$ ,  $m=1,2,\dots,5$ ,  $i=0,1,\dots,10$ , where approximations are based on  $n=25$  samples. Comparison of the approximated values with the

true values given on page A-2 shows that for  $i \geq 6$   $\hat{T}_m(\alpha_{t_i}) = T_m(\alpha_{t_i})$  for all  $m=1,2,\dots,M$ . Discrepancies for  $i < 6$  are caused by the fact that in the approximation, classes 1 and 4 were split before classes 2 and 5 and classes 2 and 3 before classes 1 and 3.

Although the approximated values  $\hat{\alpha}_{t_i}$ ,  $i=0,1,\dots,10$  do not seem close to the true values  $\alpha_{t_i}$ ,  $i=0,1,\dots,10$ , the computational efficiency  $\mathcal{CE}(\hat{\alpha}_{t_i})$  for the approximated values (page A-8) is only slightly less than the computational efficiency  $\mathcal{CE}(\alpha_{t_i})$  for the true values (page A-3). Thus if the sets  $\hat{T}_m(\hat{\alpha}_{t_i}) = T_m(\alpha_{t_i})$ ,  $m=1,2,\dots,5$ , the estimator  $\hat{R}(\hat{\alpha}_{t_i})$  is equivalent to the estimator  $\hat{R}(\alpha_{t_i})$  and is almost as efficient.

From page A-8, we see that  $\hat{\alpha}_{t_6}^* = \hat{\alpha}_{t_6}$  maximizes the approximate computational efficiency  $\hat{\mathcal{C}}\mathcal{E}$ . Since  $\hat{T}_m(\hat{\alpha}_{t_6}^*) = T_m(\alpha_{t_6}^*)$ ,  $m=1,2,\dots,5$ , and  $\mathcal{CE}(\hat{\alpha}_{t_6}^*) = 19.49 \approx \mathcal{CE}(\alpha_{t_6}^*) = 20.99$ , the approximate optimal estimator  $\hat{R}(\hat{\alpha}_{t_6}^*)$  is almost as efficient as the true optimal estimator  $\hat{R}(\alpha_{t_6}^*)$ .

On page A-9, the final estimator is compared with the posterior and error count estimators. The final estimator is formed as the posterior estimator  $\hat{R}(p)$  on  $n=25$  samples and the approximated optimal  $\hat{R}(\hat{\alpha}_{t_6}^*)$  on the remaining  $N-25$  samples. If the final estimator uses a total of  $N=400$  samples, the computational efficiency of the final estimator relative to the posterior estimator is 1.71. By using the final estimator rather than the posterior, on the average 828.93 density evaluations have been saved in obtaining an equally accurate estimate of the risk. Thus if the work involved in approximating the optimal  $\hat{\alpha}_{t_6}^*$  with  $n=25$  samples is less than the work involved in evaluating 828.93 densities, the final estimator

is preferable.

Compared to the error count estimator, if the final estimator uses  $N=400$  samples, on the average 23,887.05 density evaluations are saved. If density evaluations are costly, one would almost surely prefer the final estimator to the error count.

Next, consider the five densities described on page B-1. The values  $d_{r_i s_i}$ , their correspondence with the points  $\alpha_{t_i}$ ,  $i=0,1,\dots,8$  and  $\alpha_{q_i}$ ,  $i=0,1,\dots,3$  and the resulting sets  $T_m(\alpha_{t_i})$  and  $Q_m(\alpha_{t_i})$ ,  $m=1,2,\dots,M$ ,  $i=0,1,\dots,8$  are given on page B-2. Page B-3 lists the values  $C(\alpha_{t_i})$ ,  $V(\alpha_{t_i})$  and  $CE(\alpha_{t_i})$ ,  $i=0,1,\dots,8$ . The value  $\alpha^* = \alpha_{t_8}$  maximizes the computational efficiency  $CE(\alpha)$  and  $CE(\alpha^*) = 17.87$ . On page B-5, the optimal estimator is compared with the error count and posterior. The computational efficiency of the optimal estimator relative to the posterior is 1.95 and relative to the error count is 15.97. The number of density evaluations saved by using the optimal estimator rather than either of the existing estimators are listed on page B-5 for various sample sizes  $N_*$  for the optimal estimator.

On page B-6 are given the points  $\hat{d}_{r_i s_i}$ ,  $i=0,1,\dots,8$ , their correspondence with  $\hat{\alpha}_{t_i}$ ,  $i=0,1,\dots,8$  and  $\hat{\alpha}_{q_i}$ ,  $i=0,1,\dots,3$  and the sets  $\hat{T}_m(\hat{\alpha}_{t_i})$  and  $\hat{Q}_m(\hat{\alpha}_{t_i})$ ,  $m=1,2,\dots,5$ ,  $i=0,1,\dots,8$ . Note that  $\hat{T}_m(\hat{\alpha}_{t_i}) = T_m(\alpha_{t_i})$ ,  $m=1,2,\dots,5$ ,  $i=0,1,\dots,8$ .

On page B-7, we see that  $\hat{\alpha}^* = \hat{\alpha}_{t_8}$  maximizes the approximate computational efficiency  $\hat{CE}$ . Thus the approximate optimal estimator  $\hat{R}(\hat{\alpha}^*)$  is equivalent to the true optimal estimator  $\hat{R}(\alpha^*)$  and its computational efficiency  $CE(\hat{\alpha}^*) = 17.51$  is only slightly less than the computational efficiency

of the optimal estimator  $CE(\alpha^*) = 17.87$ .

Again, the final estimator formed by the posterior  $\hat{R}(p)$  on  $n=25$  samples and the approximate optimal  $\hat{R}(\hat{\alpha}^*)$  on the remaining  $N-25$  samples compares favorably to the error count and posterior estimators. Page B-8 makes comparisons in terms of relative computational efficiency and saved density evaluations for various sample sizes  $N$  for the final estimator. Thus if the final estimator is based on  $N=400$  samples, its computational efficiency relative to the posterior is 1.76, and by using the final estimator rather than the posterior, on the average 665 density evaluations would be saved for an equally accurate estimate of the risk. The computational efficiency of the final estimator relative to the error count estimator is 14.38 when the final estimator uses a sample size of  $N=400$ . In order to obtain the same accuracy with the error count estimator, on the average 11,707.5 more density evaluations would be required.

CHAPTER 5  
CONCLUSIONS

5.1 Summary of Results

In this thesis we have studied estimators for the Bayes risk in terms of the amount of computation they require and their accuracy. The existing estimators for the risk, namely the error count and the posterior, were shown to be inadequate computationally, thus several new estimators for the Bayes risk have been proposed. In particular, a family of estimators, indexed on a scalar parameter  $\alpha$ , was defined in such a way that estimators in the family in general required less computation than the existing estimators. The optimal estimator was chosen as that estimator in the family with maximum computational efficiency, and had the property of requiring the least amount of computation for a given accuracy.

In estimation of the Bayes risk, point evaluations of the class conditional densities are, for many problems, the single most important factor contributing to the computational effort. For this reason, the amount of computation required for a given Bayes risk estimator was defined as the number of density evaluations involved in the estimation procedure. The existing estimators, the error count estimator and the posterior estimator, require for each sample  $X_j$ ,  $j=1,2,\dots,N$ , evaluation of the class conditional density  $f_\ell(X_j)$  for each class  $\ell=1,2,\dots,M$ , a total of  $N \times M$  density evaluations. Thus when the number of classes  $M$  is large or the number of samples  $N$  is large (an accurate estimate of the risk is desired), the existing estimators were seen to be impractical from a computational aspect.

In searching for an estimator for the Bayes risk which could be computed with fewer density evaluations per sample, a general form  $\hat{R}(T)$  for Bayes risk estimators was discovered. An estimator of the form  $\hat{R}(T)$  was defined by associating with each class  $m$  some set of classes  $T_m$ . When the number of classes  $M=2$ , the class of estimators of the general form  $\hat{R}(T)$  consisted of the two existing estimators, the error count estimator and the posterior estimator. For more than two classes, the class of estimators of the general form contained several new estimators for the Bayes risk, in addition to the existing estimators.

By restricting the set of classes associated with each class  $m$  to be those classes  $T_m(\alpha)$  that are " $\alpha$ -close" to class  $m$ , an estimator  $\hat{R}(\alpha)$  of the general form was defined which in general required fewer density evaluations to compute. As the scalar parameter  $\alpha$  varied, the sets of classes  $T_1(\alpha), \dots, T_M(\alpha)$  varied and a family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  of Bayes risk estimators was achieved. Estimators in the family were characterized by the average number of conditional density evaluations needed to compute them and by their variance.

The optimal estimator  $\hat{R}(\alpha^*)$  from the family was defined as that estimator with maximum computational efficiency, where the computational efficiency of an estimator was defined as the inverse of the product of its variance and the average number of density evaluations it required. It was shown that the optimal estimator  $\hat{R}(\alpha^*)$  required the least amount of computation to achieve a given accuracy, or, symmetrically, achieved the greatest accuracy for a fixed amount of computation.

It was pointed out that in practice, the optimal estimator  $\hat{R}(\alpha^*)$  could not be determined by maximizing the computational efficiency,

since this in effect would require knowledge of the true risk  $R$ . Thus a method was proposed whereby a subset  $n$  of the total  $N$  samples was used to approximate the optimal estimator. The  $n$  samples should contain enough information on the closeness of the classes to determine an almost optimal estimator. The remaining  $N-n$  samples would be used in the approximate optimal estimator to obtain an accurate estimate of the risk with a minimum of computation.

For both examples given in the appendix, the optimal estimator was closely approximated using  $n=25$  samples. In fact, for each case the approximate optimal estimator was equivalent to the true optimal estimator, in the sense that point estimates of the risk resulting from either would be identical. However, the approximate optimal estimator would, on the average, perform slightly more density evaluations in forming the estimate.

The technique for approximating the optimal estimator forms as a by product the posterior estimator based  $n$  samples. Because the posterior estimator has minimum variance among all estimators considered here, we defined as our final estimator the posterior estimator on the  $n$  samples used in approximating the optimal estimator and the approximated optimal estimator on the remaining  $N-n$  samples.

The final estimator was compared to the error count and posterior estimators. The comparisons were based on the number of density evaluations that would be saved by using the final estimator rather than one of the existing estimators in obtaining equally accurate estimates of the risk. Situations for which great computational savings would be expected were the following:

1. an accurate estimate of the risk is desired, thus a large number  $N$  of samples is used.
2. the number of classes  $M$  is large and the classes tend to form several small clusters.

### 5.2 Recommendations for Further Research

In section 3.2.5, variances for estimators in the family  $\{\hat{R}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  based on unrestricted sampling were derived. We discussed properties of these variances and indicated reasons for believing that as  $\alpha$  increased, the variance of the estimator  $\hat{R}(\alpha)$  should not decrease. Thus the question: under what conditions does the following proposition hold?

#### Proposition 1.

If  $0 \leq \alpha_1 \leq \alpha_2 < \alpha_{\max}$   
 then  $V(\alpha_1) \leq V(\alpha_2)$ .

Aside from a theoretical interest, proposition 1 has the practical consequence indicated in corollary 4-2. That was that the computational efficiency could be maximized over the  $J+1$  values  $\alpha_{q_0} \dots \alpha_{q_J}$  rather than over the larger number  $K+1$  of values  $\alpha_{t_0} \dots \alpha_{t_K}$ . Thus to determine the optimal estimator for example 1 in the appendix, the computational efficiency need only be evaluated at  $J + 1 = 6$  points rather than  $K + 1 = 11$  points, and at  $J + 1 = 4$  points rather than  $K + 1 = 9$  points for example 2.

The same question of non-decreasing variances arises in the family  $\{\hat{SR}(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  of risk estimators based on stratified sampling. Moore, Whitsitt and Landgrebe's example [30], in which the stratified

error count estimator had smaller variance than the stratified posterior estimator, shows that an analog to theorem 3-1 is not possible for stratified sampling. However, for their example, the error count estimator would not be included in the family. Thus one might still hope to show that the variances of estimators in the family  $\{\hat{S}R(\alpha) : 0 \leq \alpha < \alpha_{\max}\}$  are non-decreasing as  $\alpha$  increases.

Finally, concerning the choice of the number of samples  $n_i$  from each class  $i=1, \dots, M$  to be used in the stratified estimator  $\hat{S}R(\alpha)$ . We discussed two choices, the heuristic one with  $n_i$  proportional to the prior probability of class  $i$ , and the choice  $n_i^*$  to minimize the variance of the estimator  $\hat{S}R(\alpha)$ . In view of chapter 4, a better choice would have been to choose  $n_i^{**}$  to maximize the computational efficiency of the estimator  $\hat{S}R(\alpha)$ . What implications would this have, in terms of the optimal estimator for stratified sampling, and could this be useful in practice?

## BIBLIOGRAPHY

1. Bahl, L.R. and Jelinek, F., "Decoding for Channels with Insertions, Deletions and Substitutions with Applications to Speech Recognition", IEEE Trans. Info. Theory, IT-21, No. 4, pp. 404-411, July, 1975.
2. Chow, C.K., "An Optimum Character Recognition System Using Decision Functions", IRE Trans. on Elec. Comp. EC-6, pp. 247-254, Dec. 1957.
3. Chow, C.K., "On Optimum Recognition Error and Reject Tradeoff" IEEE Trans. Info. Theory, IT-16, pp. 41-46, Jan. 1970.
4. Cover, T.M. and Wagner, T.J., "Topics in Statistical Pattern Recognition" in Digital Patterns Recognition, K.S. Fu, Ed., New York: Springer-Verlag Publishers, 1976.
5. de Figueiredo, R.J.P., Pau, K.C., Sagar, A.O., Starks, S.A., and VanRooy, E.C., "An Algorithm for Extraction of More Than One Optimal Linear Feature from Several Gaussian Pattern Classes", ICSA Technical Report #275-025-026, ICSA, Rice University, Houston, Texas, April 1976.
6. Duda, R.O. and Hart, P.E., Pattern Classification and Scene Analysis, New York: Wiley, 1973.
7. Ferguson, T.S., Mathematical Statistics - A Decision Theoretic Approach, New York: Academic Press, 1969.
8. Fisher, F.P., and Patrick, E.A., "A Preprocessing Algorithm for Nearest Neighbor Decision Rules", Proc. Nat. Electronics Conf., vol 26, pp. 481-485, Dec. 1970.
9. Fralick, S.C. and Scott, R.W., "Non-Parametric Bayes Risk Estimation", IEEE Trans. Info. Theory, IT-17, pp. 440-444, July 1971.
10. Friedman, J.H., Baskett, F. and Shustek, L.S., "An Algorithm for Finding Nearest Neighbors", IEEE Trans. Comp. C-24, pp. 1000-1006, Oct. 1975.
11. Fukunaga, K. and Kessel, D., "Application of Optimum Error Reject Functions", IEEE Trans. Info. Theory, IT-18, pp. 814-817, Nov. 1972.
12. Fukunaga, K. and Kessel, D., "Estimation of Classification Error", IEEE Trans. Comp., C-20, pp. 1521-1527, Dec. 1971.
13. Fukunaga, K. and Hostetter, L.D., "k-Nearest Neighbor Bayes Risk Estimation", IEEE Trans. Info. Theory, IT-21, pp. 285-293, May 1975.
14. Fukunaga, K. and Narendra, P.M., "A Branch and Bound Algorithm for Computing k-Nearest Neighbors", IEEE Trans. Comp., C-24, No. 7, pp. 750-753, July 1975.

15. Fukunaga, K. and Kessel, D.L., "Non-Parametric Bayes Error Estimation Using Unclassified Samples", IEEE Trans. Info. Theory, IT-19, pp. 430-440, July 1973.
16. Hammersly, J.M. and Handscomb, D.C., Monte Carlo Methods, London: Methuen and Co., 1964.
17. Hammersly, J.M. and Morton, K.W., "A New Monte Carlo Technique: Varieties", Proceedings of the Cambridge Philosophical Society, Vol. 52, pp. 449-475, July 1956.
18. Hart, P.E., "Condensed Nearest Neighbor Rule", IEEE Trans. Info. Theory, IT-14, pp. 515-516, May 1968.
19. Highleyman, W.H., "The Design and Analysis of Pattern Recognition Experiments", Bell Systems Technical Journal, Vol. 41, pp. 723-744, 1962.
20. Hills, M., "Allocation Rules and Their Error Rates" J. Roy. Statis. Soc., Series B, No. 28, pp. 1-31, 1966.
21. Hoel, Port, Stone, Introduction to Probability Theory, Boston: Houghton Mifflin Co., 1971.
22. Hoel, Port, Stone, Introduction to Statistical Theory, Boston: Houghton Mifflin Co., 1971.
23. Jelinek, F., Bahl, L.R., Mercer, R.L., "Design of a Linguistic Statistical Decoder for Recognition of Continuous Speech", IEEE Trans. Info. Theory, IT-21, pp. 250-256, May 1976.
24. Kanal, L., "Patterns in Pattern Recognition: 1968-1974, IEEE Trans. Info. Theory, IT-20, No. 6, Nov. 1974.
25. Kanal, L., and Chandrasekaran, B., "On Dimensionality and Sample Size in Statistical Pattern Classification", Proc. Nat. Electron. Conf., Vol. XXIV, pp. 2-7, Dec. 9-11, 1968.
26. Lachenbruch, P.A. and Mickey, M.R., "Estimation of Error Rates in Discriminant Analysis", Technometrics, Vol. 10, pp. 1-11, Feb. 1968.
27. Lissack, T. and Fu, K.S., "Error Estimation in Pattern Recognition via  $L^2$  Distance Between Posterior Density Functions", IEEE Trans. Info. Theory, IT-22, No. 1, pp. 34-45, Jan. 1976.
28. Lissack, T. and Fu, K.S., "Parametric Feature Extraction Through Error Minimization Applied to Medical Diagnosis", IEEE Trans. Sys. Man. Cyber, SMC-6, No. 9, Sept. 1976.
29. Loftsgaarden, D.O. and Quesenberry, C.P., "A Non-Parametric Estimate of a Multivariate Density Function", Ann. Math. Stat. Vol. 36,

pp. 1049-1051, 1965.

30. Moore, D.S., Whitsitt, S.J. and Landgrebe, D.A., "Variance Comparisons for Unbiased Estimators of Probability of Correct Classification", IEEE Trans. Info. Theory, IT-22, No. 1, pp. 102-105, Jan. 1976.
31. Neyman, Jerzy, "On the Two Different Aspects of the Representative Method. The Method of Stratified Sampling and the Method of Purposive Selection", Journal of the Royal Statistical Society", Vol. 97, pp. 558-625, 1934.
32. Parzen, E., "On Estimation of a Probability Density Function and Mode", Ann. Math. Stat., 33, pp. 1065-1076, Sept. 1962.
33. Pfeiffer, P. and Schum, D., Introduction to Applied Probability, New York: Academic Press, 1973.
34. Rao, C.R., Linear Statistical Inference and Its Applications, New York: John Wiley and Sons, Inc. 1965.
35. Starks, S.A., de Figueiredo, R.J.P., and VanRooy, D.L., "An Algorithm for Optimal Single Linear Feature Extraction From Several Gaussian Pattern Classes", ICSA Tech. Report #275-025-022, ICSA, Rice University, Houston, Texas, Nov. 1975
36. Swonger, C.W., "Sample Set Condensation for a Condensed Nearest Neighbor Decision Rule For Pattern Recognition", in Frontiers of Pattern Recognition, S. Watanabe, Ed., New York: Academic Press, 1972.
37. Toussaint, P.T., "Bibliography on Estimation of Misclassification", IEEE Trans. Info. Theory, IT-20, pp. 472-479, July 1974.
38. Wagner, T.J., "Deleted Estimates of the Bayes Risk", Annals of Statistics, Vol. 1, No. 2, pp. 359-362, March, 1973.
39. Whitsitt, S.J. and Landgrebe, D.A., "Error Estimation and Separability Measures in Feature Selection for Multiclass Pattern Recognition", Laboratory for Applications of Remote Sensing Report No. 082377, W. Lafayette, Indiana, Aug. 1977.
40. Yunk, T.P., "A Technique to Identity Nearest Neighbors", IEEE Trans. Sys. Man. Cyber., SMC-6, No. 10, pp. 678-683, Oct. 1976.

AD-A055 997

RICE UNIV HOUSTON TEX DEPT OF ELECTRICAL ENGINEERING  
COMPUTATIONALLY EFFICIENT ESTIMATORS FOR THE BAYES RISK.(U)  
MAY 78 L D WILCOX, R J FIGUEIREDO

F/G 5/8

AFOSR-75-2777

UNCLASSIFIED

EE-TR-7804

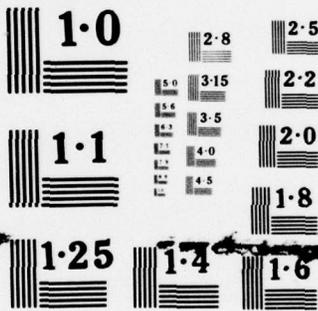
AFOSR-TR-78-1081

NL

2 OF 2  
ADA  
065997



END  
DATE  
FILMED  
8-78  
DDC



NATIONAL BUREAU OF STANDARDS  
MICROCOPY RESOLUTION TEST CHART

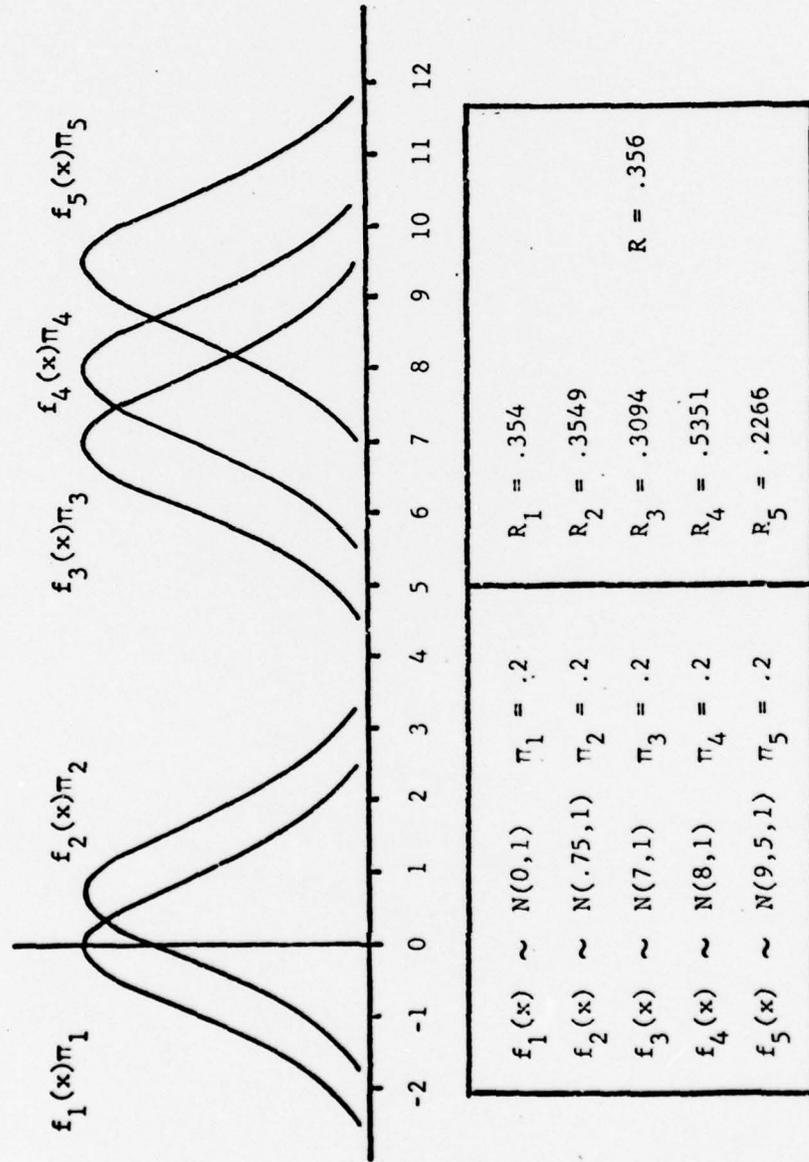
#### ACKNOWLEDGMENTS

Thanks to Dr. Fred Jelinek for the initial suggestions which led to this research, and to Dr. Paul Pfeiffer and Dr. Terry Wagner for many helpful discussions. Also thanks to Mr. Sergio Batiz for drawing the pictures and to Jeanne Fulton for so patiently typing all this.

This work was supported by AFOSR Grant No. 75-2777.

APPENDIX

A. Data From Example 1.



Example 1. Values of  $\alpha_{t_i}$ ,  $\alpha_{q_i}$ ,  $d_{r_i s_i}$ ,  $T_m(\alpha_{t_i})$  and  $Q_m(\alpha_{t_i})$

$\alpha$	$d_{r_i s_i}$	$\alpha_{t_i}$	$\alpha_{q_i}$	$T_1(\alpha)$	$T_2(\alpha)$	$T_3(\alpha)$	$T_4(\alpha)$	$T_5(\alpha)$	$Q_1(\alpha)$	$Q_2(\alpha)$	$Q_3(\alpha)$	$Q_4(\alpha)$	$Q_5(\alpha)$
0	-	$\alpha_{t_0}$	$\alpha_{q_0}$	12345	12345	12345	12345	12345	12345	12345	12345	12345	12345
.000001	$d_{15}$	$\alpha_{t_1}$		1234	12345	12345	12345	2345	12345	12345	12345	12345	12345
.000005	$d_{25}$	$\alpha_{t_2}$		1234	1234	12345	12345	345	12345	12345	12345	12345	12345
.000027	$d_{14}$	$\alpha_{t_3}$		123	1234	12345	2345	345	12345	12345	12345	12345	12345
.000122	$d_{24}$	$\alpha_{t_4}$		123	123	12345	345	345	12345	12345	12345	12345	12345
.000174	$d_{13}$	$\alpha_{t_5}$	$\alpha_{q_1}$	12	123	2345	345	345	123	12345	12345	2345	2345
.00062	$d_{23}$	$\alpha_{t_6}$	$\alpha_{q_2}$	12	12	345	345	345	12	12	345	345	345
.03652	$d_{35}$	$\alpha_{t_7}$		12	12	34	345	45	12	12	345	345	345
.06022	$d_{45}$	$\alpha_{t_8}$	$\alpha_{q_3}$	12	12	34	34	5	12	12	34	34	5
.07042	$d_{34}$	$\alpha_{t_9}$	$\alpha_{q_4}$	12	12	3	4	5	12	12	3	4	5
.0744	$d_{12}$	$\alpha_{t_{10}}$	$\alpha_{q_5}$	1	2	3	4	5	1	2	3	4	5

Example 1. Values of  $C(\alpha_{t_i})$ ,  $V^*(\alpha_{t_i})$  and  $CE(\alpha_{t_i})$

$\alpha$	$\alpha_{t_i}$	$\alpha_{q_i}$	$C(\alpha)$	$V^*(\alpha)$	$CE(\alpha)$
0	$\alpha_{t_0}$	$\alpha_{q_0}$	5.00	.01831	10.92
.000001	$\alpha_{t_1}$		5.00	.01831	10.92
.000005	$\alpha_{t_2}$		5.0	.01831	10.92
.000027	$\alpha_{t_3}$		5.0	.01831	10.92
.000122	$\alpha_{t_4}$		5.0	.01831	10.92
.000174	$\alpha_{t_5}$	$\alpha_{q_1}$	4.2	.01831	13.00
.00062	$\alpha_{t_6}$	$\alpha_{q_2}$	2.6	.01832	20.99
.03652	$\alpha_{t_7}$		3.11	.02425	13.26
.06022	$\alpha_{t_8}$	$\alpha_{q_3}$	3.25	.07588	4.05
.07042	$\alpha_{t_9}$	$\alpha_{q_4}$	3.62	.14358	1.92
.0744	$\alpha_{t_{10}}$	$\alpha_{q_5}$	3.83	.22984**	1.14

\*Estimates based on 600 samples

\*\*Analytic value is .22926

Example 1. Matrix of covariances  $[C_{m\ell}(\alpha)]^*$  for  $\alpha = \alpha_{t_0}, \alpha_{t_8}, \alpha_{t_{10}}$ .

$[C_{m\ell}(\alpha)]^*$

$\alpha_{t_0}$ (posterior)	.582	-.127	-.122	-.215	-.077
		.520	-.103	-.181	-.165
			.35	-.088	.166
				.588	-.071
					.185
$\alpha_{t_8}$	.582	-.127	-.124	-.222	-.184
		.520	-.105	-.187	-.071
			.456	-.049	-.069
				1.429	-.124
					1.038
$\alpha_{t_{10}}$ (error count)	1.843	-.107	-.114	-.247	-.087
		1.265	-.076	-.165	-.058
			1.339	-.175	-.062
				2.708	-.134
					1.038
$\alpha_{t_{10}}$ (analytic)	1.644	-.126	-.110	-.189	-.080
		1.644	-.110	-.190	-.080
			1.45	-.166	-.070
				2.39	-.121
					1.08

\*Estimates based on 600 samples.

Example 1. Estimators  $\hat{R}(\alpha_{t_i})$   $i=0,1,\dots,10$  For Various Sample Sizes

$\alpha$	$\alpha_{t_i}$	$\alpha_{q_i}$	N=50 $\hat{R}(\alpha)$	N=100 $\hat{R}(\alpha)$	N=200 $\hat{R}(\alpha)$	true R
* 0	$\alpha_{t_0}$	$\alpha_{q_0}$	.36047	.35832	.36322	.356
.000001	$\alpha_{t_1}$		.36047	.35832	.36322	.356
.000005	$\alpha_{t_2}$		.36047	.35832	.36322	.356
.000027	$\alpha_{t_3}$		.36047	.35832	.36322	.356
.000122	$\alpha_{t_4}$		.36047	.35832	.36322	.356
.000174	$\alpha_{t_5}$	$\alpha_{q_1}$	.36047	.35833	.36323	.356
.00062	$\alpha_{t_5}$	$\alpha_{q_2}$	.36045	.35829	.36318	.356
.03652	$\alpha_{t_7}$		.37872	.36578	.37265	.356
.06022	$\alpha_{t_8}$	$\alpha_{q_3}$	.33586	.38022	.37872	.356
.07042	$\alpha_{t_9}$	$\alpha_{q_4}$	.34302	.37835	.38116	.356
** .0744	$\alpha_{t_{10}}$	$\alpha_{q_5}$	.28	.35	.365	.356

\*  $\hat{R}(\alpha_{t_0})$  is equivalent to the posterior estimator  $\hat{R}(p)$ .

\*\*  $\hat{R}(\alpha_{t_{10}})$  equivalent to the error count estimator  $\hat{R}(ec)$ .

## Example 1

Computational efficiency of the optimal estimator relative to the posterior estimator and the number of density evaluations saved by using the optimal estimator rather than the posterior, for various sample sizes  $N_*$  for the optimal estimator.

$N_*$	$RCE(\alpha^*, p)$	$S(\alpha^*, p)$
100	1.92	239
200	1.92	478
400	1.92	956
600	1.92	1,434

Computational efficiency of the optimal estimator relative to the error count estimator and the number of density evaluations saved by using the optimal estimator rather than the error count, for various sample sizes  $N_*$  for the optimal estimator.

$N_*$	$RCE(\alpha^*, ec)$	$S(\alpha^*, ec)$
100	24.07	5,998
200	24.07	11,996
400	24.07	23,992
600	24.07	35,988

Example 1. Values of  $\hat{\alpha}_{t_i}$ ,  $\hat{\alpha}_{q_i}$ ,  $\hat{d}_{r_i s_i}$ ,  $\hat{t}_m(\hat{\alpha}_{t_i})$  and  $\hat{Q}_m(\hat{\alpha}_{t_i})$  Approximated On the Basis of n=25 Samples.

$\hat{\alpha}$	$\hat{d}_{r_i s_i}$	$\hat{\alpha}_{t_i}$	$\hat{\alpha}_{q_i}$	$\hat{t}_1(\hat{\alpha})$	$\hat{t}_2(\hat{\alpha})$	$\hat{t}_3(\hat{\alpha})$	$\hat{t}_4(\hat{\alpha})$	$\hat{t}_5(\hat{\alpha})$	$\hat{Q}_1(\hat{\alpha})$	$\hat{Q}_2(\hat{\alpha})$	$\hat{Q}_3(\hat{\alpha})$	$\hat{Q}_4(\hat{\alpha})$	$\hat{Q}_5(\hat{\alpha})$
0	-	$\hat{\alpha}_{t_0}$	$\hat{\alpha}_{q_0}$	12345	12345	12345	12345	12345	12345	12345	12345	12345	12345
.000306	$\hat{d}_{15}$	$\hat{\alpha}_{t_1}$		1234	12345	12345	12345	2345	12345	12345	12345	12345	12345
.014804	$\hat{d}_{14}$	$\hat{\alpha}_{t_2}$		123	12345	12345	2345	2345	12345	12345	12345	12345	12345
.0003063	$\hat{d}_{25}$	$\hat{\alpha}_{t_3}$		123	1234	12345	2345	345	12345	12345	12345	12345	12345
.014804	$\hat{d}_{24}$	$\hat{\alpha}_{t_4}$		123	123	12345	345	345	12345	12345	12345	12345	12345
.049186	$\hat{d}_{23}$	$\hat{\alpha}_{t_5}$	$\hat{\alpha}_{q_1}$	123	12	1345	345	345	12345	123	12345	1345	1345
.017754	$\hat{d}_{13}$	$\hat{\alpha}_{t_6}$	$\hat{\alpha}_{q_2}$	12	12	345	345	345	12	12	345	345	345
.044470	$\hat{d}_{35}$	$\hat{\alpha}_{t_7}$		12	12	34	345	45	12	12	345	345	345
.063190	$\hat{d}_{45}$	$\hat{\alpha}_{t_8}$	$\hat{\alpha}_{q_3}$	12	12	34	34	5	12	12	34	34	5
.073341	$\hat{d}_{34}$	$\hat{\alpha}_{t_9}$	$\hat{\alpha}_{q_4}$	12	12	3	4	5	12	12	3	4	5
.074388	$\hat{d}_{12}$	$\hat{\alpha}_{5_{10}}$	$\hat{\alpha}_{q_5}$	1	2	3	4	5	1	2	3	4	5

Example 1. Values of  $\hat{C}(\hat{\alpha}_{t_i})$ ,  $\hat{V}(\hat{\alpha}_{t_i})$  and  $\hat{CE}(\hat{\alpha}_{t_i})$  Approximated On the Basis of  $n=25$  Samples. Also  $CE(\hat{\alpha}_{t_i})$ .

$\hat{\alpha}$	$\hat{\alpha}_{t_i}$	$\hat{\alpha}_{q_i}$	$\hat{C}(\hat{\alpha})$	$\hat{V}(\hat{\alpha})$	$\hat{CE}(\hat{\alpha})$	$CE(\hat{\alpha})$
0	$\hat{\alpha}_{t_0}$	$\hat{\alpha}_{q_0}$	5.	.01973	10.14	10.92
.000306	$\hat{\alpha}_{t_1}$		5.	.01973	10.14	10.92
.014804	$\hat{\alpha}_{t_2}$		5.	.01973	10.14	10.92
.0003063	$\hat{\alpha}_{t_3}$		5.	.01973	10.14	10.92
.014804	$\hat{\alpha}_{t_4}$		5.	.01973	10.14	10.92
.049186	$\hat{\alpha}_{t_5}$	$\hat{\alpha}_{q_1}$	4.48	.01973	11.31	12.24
.017754	$\hat{\alpha}_{t_6}$	$\hat{\alpha}_{q_2}$	2.78	.01973	18.23	19.49
.04447	$\hat{\alpha}_{t_7}$		3.19	.03124	10.03	12.61
.06319	$\hat{\alpha}_{t_8}$	$\hat{\alpha}_{q_3}$	3.56	.06121	4.59	3.88
.073341	$\hat{\alpha}_{t_9}$	$\hat{\alpha}_{q_4}$	3.96	.11432	2.23	1.81
.074388	$\hat{\alpha}_{t_{10}}$	$\hat{\alpha}_{q_5}$	4.26	.16667	1.41	1.14

## Example 1.

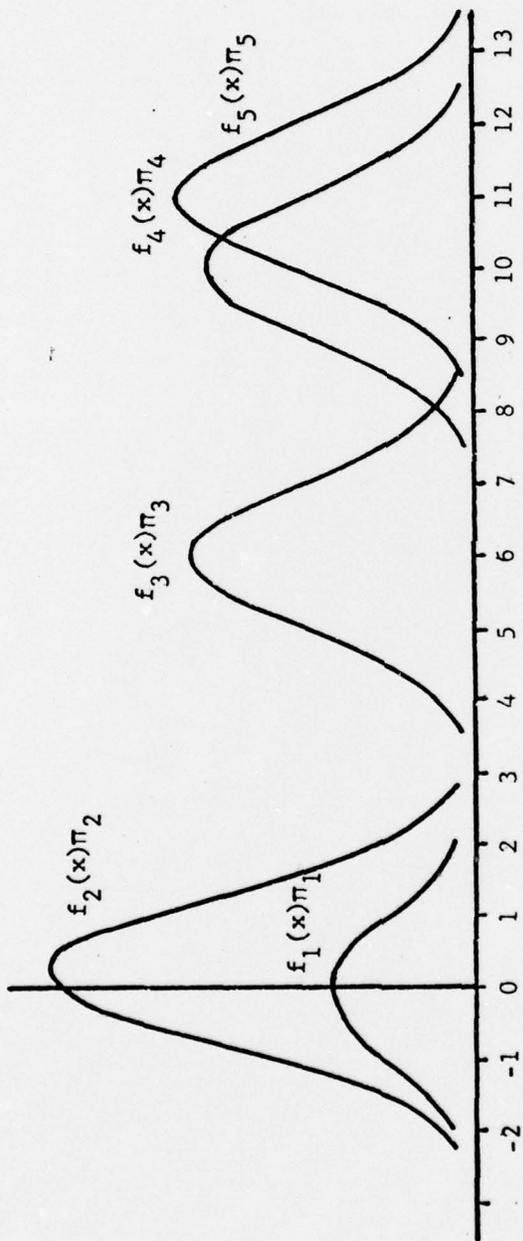
Computational efficiency of the final estimator relative to the posterior estimator and the number of density evaluations saved by using the final estimator rather than the posterior, when the optimal estimator is approximated on the basis of  $n=25$  samples, for various sample sizes  $N$  for the final estimator.

$N$	$\rho CE(f,p)$	$\$ (f,)$
100	1.5	166.75
200	1.64	391.36
400	1.71	828.93
600	1.74	1,275.39

Computational efficiency of the final estimator relative to the error count estimator and the number of density evaluations saved by using the final estimator rather than the error count, when the optimal estimator is approximated on the basis of  $n=25$  samples, for various sample sizes  $N$  for the final estimator.

$N$	$\rho CE(f,ec)$	$\$ (f,ec)$
100	18.78	5,929.63
200	20.49	11,918.14
400	21.46	23,887.05
600	21.81	35,866.04

B. Data From Example 2.



$f_1(x) \sim N(0,1)$	$\pi_1 = .1$	$R_1 = .9738$
$f_2(x) \sim N(.5,1)$	$\pi_2 = .3$	$R_2 = .0097$
$f_3(x) \sim N(6,1)$	$\pi_3 = .2$	$R_3 = .026$
$f_4(x) \sim N(10,1)$	$\pi_4 = .19$	$R_4 = .3685$
$f_5(x) \sim N(11,1)$	$\pi_5 = .21$	$R_5 = .2743$
		$R = .233$

Example 2: Values of  $\alpha_{t_i}$ ,  $\alpha_{q_i}$ ,  $d_{r_i s_i}$ ,  $T_1(\alpha)$ ,  $T_2(\alpha)$ ,  $T_3(\alpha)$ ,  $T_4(\alpha)$ ,  $T_5(\alpha)$ ,  $Q_1(\alpha)$ ,  $Q_2(\alpha)$ ,  $Q_3(\alpha)$ ,  $Q_4(\alpha)$  and  $Q_5(\alpha)$ .

$\alpha$	$d_{r_i s_i}$	$\alpha_{t_i}$	$\alpha_{q_i}$	$T_1(\alpha)$	$T_2(\alpha)$	$T_3(\alpha)$	$T_4(\alpha)$	$T_5(\alpha)$	$Q_1(\alpha)$	$Q_2(\alpha)$	$Q_3(\alpha)$	$Q_4(\alpha)$	$Q_5(\alpha)$
0	-	$\alpha_{t_0}$	$\alpha_{q_0}$	12345	12345	12345	12345	12345	12345	12345	12345	12345	12345
.000000021	$d_{15}$	$\alpha_{t_1}$		1234	12345	12345	12345	2345	12345	12345	12345	12345	12345
.0000000105	$d_{25}$	$\alpha_{t_2}$		1234	1234	12345	12345	345	12345	12345	12345	12345	12345
.00000002	$d_{14}$	$\alpha_{t_3}$		123	1234	12345	2345	345	12345	12345	12345	12345	12345
.00000012	$d_{24}$	$\alpha_{t_4}$		123	123	12345	345	345	12345	12345	12345	12345	12345
.0006226	$d_{13}$	$\alpha_{t_5}$	$\alpha_{q_1}$	12	123	2345	345	345	123	12345	12345	2345	2345
.0022214	$d_{23}$	$\alpha_{t_6}$	$\alpha_{q_2}$	12	12	345	345	345	12	12	345	345	345
.003592	$d_{35}$	$\alpha_{t_7}$		12	12	34	345	45	12	12	345	345	345
.0105239	$d_{34}$	$\alpha_{t_8}$	$\alpha_{q_3}$	12	12	3	45	45	12	12	3	45	45

Example 2. Values of  $C(\alpha_{t_i})$ ,  $V^*(\alpha_{t_i})$  and  $CE(\alpha_{t_i})$

$\alpha$	$\alpha_{t_i}$	$\alpha_{q_i}$	$C(\alpha)$	$V^*(\alpha)$	$CE(\alpha)$
0	$\alpha_{t_0}$	$\alpha_{q_0}$	5	.02184	9.16
.000000021	$\alpha_{t_1}$		5	.02184	9.16
.000000105	$\alpha_{t_2}$		5	.02184	9.16
.0000002	$\alpha_{t_3}$		5	.02184	9.16
.0000012	$\alpha_{t_4}$		5	.02184	9.16
.0006226	$\alpha_{t_5}$	$\alpha_{q_1}$	4.4	.02184	10.41
.0022214	$\alpha_{t_6}$	$\alpha_{q_2}$	2.62	.02205	17.31
.003592	$\alpha_{t_7}$		2.66	.0224	16.78
.0105237	$\alpha_{t_8}$	$\alpha_{q_3}$	1.94	.02884	17.87

\*Estimates based on 600 samples.

Example 2. Estimators  $\hat{R}(\alpha_{t_i})$   $i=0,1,\dots,8$  and  $\hat{R}(ec)$  for Various Sample Sizes.

$\alpha$	$\alpha_{t_i}$	$\alpha_{q_i}$	N=50 $\hat{R}(\alpha)$	N=100 $\hat{R}(\alpha)$	N=200 $\hat{R}(\alpha)$	true R
* 0	$\alpha_{t_0}$	$\alpha_{q_0}$	.21929	.22522	.22067	.233
.000000021	$\alpha_{t_1}$		.21929	.22522	.22067	.233
.000000105	$\alpha_{t_2}$		.21929	.22522	.22067	.233
.0000002	$\alpha_{t_3}$		.21929	.22522	.22067	.233
.0000012	$\alpha_{t_4}$		.21929	.22522	.22067	.233
.0006226	$\alpha_{t_5}$	$\alpha_{q_1}$	.21932	.22537	.22078	.233
.00222214	$\alpha_{t_6}$	$\alpha_{q_2}$	.21911	.22414	.21977	.233
.003592	$\alpha_{t_7}$		.21915	.22590	.22093	.233
.0105237	$\alpha_{t_8}$	$\alpha_{q_3}$	.19653	.21974	.21861	.233
** ec	-	-	.16	.22	.2	

\* $\hat{R}(\alpha_{t_0})$  is equivalent to the posterior estimator  $\hat{R}(p)$ .

\*\* $\hat{R}(ec)$  is not allowed in the family.

## Example 2.

Computational efficiency of the optimal estimator relative to the posterior estimator and the number of density evaluations saved by using the optimal estimator rather than the posterior for various sample sizes  $N_*$  for the optimal estimator.

$N_*$	$RCE(\alpha^*, p)$	$S(\alpha^*, p)$
100	1.95	184
200	1.95	368
400	1.95	736
600	1.95	1,104

Computational efficiency of the optimal estimator relative to the error count estimator and the number of density evaluations saved by using the optimal estimator rather than the error count, for various sample sizes  $N_*$  for the optimal estimator.

$N_*$	$RCE(\alpha^*, ec)$	$S(\alpha^*, ec)$
100	15.97	2,904
200	15.97	5,808
400	15.97	11,616
600	15.97	17,424

Example 2. Values of  $\hat{\alpha}_{t_i}$ ,  $\hat{\alpha}_{q_i}$ ,  $\hat{d}_{r_i s_i}$ ,  $\hat{r}_{i s_i}$ ,  $\hat{m}(\hat{\alpha}_{t_i})$  and  $\hat{Q}_m(\hat{\alpha}_{t_i})$  Approximated On the Basis of n=25 Samples

$\hat{\alpha}$     $\hat{d}_{r_i s_i}$     $\hat{\alpha}_{t_i}$     $\hat{\alpha}_{q_i}$     $\hat{r}_1(\hat{\alpha})$     $\hat{r}_2(\hat{\alpha})$     $\hat{r}_3(\hat{\alpha})$     $\hat{r}_4(\hat{\alpha})$     $\hat{r}_5(\hat{\alpha})$     $\hat{Q}_1(\hat{\alpha})$     $\hat{Q}_2(\hat{\alpha})$     $\hat{Q}_3(\hat{\alpha})$     $\hat{Q}_4(\hat{\alpha})$     $\hat{Q}_5(\hat{\alpha})$

0	-	$\hat{\alpha}_{t_0}$	$\hat{\alpha}_{q_0}$	12345	12345	12345	12345	12345	12345	12345	12345	12345	12345	12345
0	$\hat{d}_{15}$	$\hat{\alpha}_{t_1}$		1234	12345	12345	12345	2345	12345	12345	12345	12345	12345	12345
.0000001	$\hat{d}_{25}$	$\hat{\alpha}_{t_2}$		1234	1234	12345	12345	345	12345	12345	12345	12345	12345	12345
.0000002	$\hat{d}_{14}$	$\hat{\alpha}_{t_3}$		123	1234	12345	2345	345	12345	12345	12345	12345	12345	12345
.00000023	$\hat{d}_{24}$	$\hat{\alpha}_{t_4}$		123	123	12345	345	345	12345	12345	12345	12345	12345	12345
.0157481	$\hat{d}_{13}$	$\hat{\alpha}_{t_5}$	$\hat{\alpha}_{q_1}$	12	123	2345	345	345	123	12345	12345	2345	2345	2345
.0447042	$\hat{d}_{23}$	$\hat{\alpha}_{t_6}$	$\hat{\alpha}_{q_2}$	12	12	345	345	345	12	12	345	345	345	345
.0124001	$\hat{d}_{35}$	$\hat{\alpha}_{t_7}$		12	12	34	345	45	12	12	345	345	345	345
.0124051	$\hat{d}_{34}$	$\hat{\alpha}_{t_8}$	$\hat{\alpha}_{q_3}$	12	12	3	45	45	12	12	3	45	45	45

Example 2. Values of  $\hat{C}(\hat{\alpha}_{t_i})$ ,  $\hat{V}(\hat{\alpha}_{t_i})$  and  $\mathcal{CE}(\hat{\alpha}_{t_i})$  Approximated On the Basis of  $n=25$  samples. Also  $\mathcal{CE}(\hat{\alpha}_{t_i})$

$\hat{\alpha}$	$\hat{\alpha}_{t_i}$	$\hat{\alpha}_{q_i}$	$\hat{C}(\hat{\alpha})$	$\hat{V}(\hat{\alpha})$	$\mathcal{CE}(\hat{\alpha})$	$\mathcal{CE}(\hat{\alpha})$
.0	$\hat{\alpha}_{t_0}$	$\hat{\alpha}_{q_0}$	5.	.01994	10.03	9.16
.0	$\hat{\alpha}_{t_1}$		5.	.01994	10.03	9.16
.0000001	$\hat{\alpha}_{t_2}$		5.	.01994	10.03	9.16
.0000002	$\hat{\alpha}_{t_3}$		5.	.01994	10.03	9.16
.0000023	$\hat{\alpha}_{t_4}$		5.	.01994	10.03	9.16
.0157484	$\hat{\alpha}_{t_5}$	$\hat{\alpha}_{q_1}$	4.46	.01994	11.24	10.29
.0447042	$\hat{\alpha}_{t_6}$	$\hat{\alpha}_{q_2}$	3.24	.01995	15.47	13.22
.0124001	$\hat{\alpha}_{t_7}$		2.72	.01921	19.14	16.35
.0124051	$\hat{\alpha}_{t_8}$	$\hat{\alpha}_{q_3}$	2.00	.02068	24.18	17.51

## Example 2.

Computational efficiency of the final estimator relative to the posterior estimator and the number of density evaluations saved by using the final estimator rather than the posterior, when the optimal estimator is approximated on the basis of  $n=25$  samples, for various sample sizes  $N$  for the final estimator.

$N$	$RCE(f,p)$	$S(f,p)$
100	1.46	126.5
200	1.64	304
400	1.76	665
600	1.8	1,020

Computational efficiency of the final estimator relative to the error count estimator and the number of density evaluations saved by using the final estimator rather than the error count, when the optimal estimator is approximated on the basis of  $n=25$  samples, for various sample sizes  $N$  for the final estimator.

$N$	$RCE(f,ec)$	$S(f,ec)$
100	11.99	3022.25
200	13.45	5913.75
400	14.38	11707.5
600	14.73	17505.75