INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in “sectioning” the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from “photographs” if essential to the understanding of the dissertation. Silver prints of “photographs” may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

University Microfilms International
300 North Zeib Road
Ann Arbor, Michigan 48106 USA
St. John's Road, Tyler's Green
High Wycombe, Bucks, England HP10 8HR
LI, Robert Yuan-Shih, 1946-
SEQUENTIAL PROCEDURES AND DISTANCE
MEASURES IN DATA PROCESSING.

Rice University, Ph.D., 1977
Engineering, electronics and electrical

Xerox University Microfilms, Ann Arbor, Michigan 48106
RICE UNIVERSITY

Sequential Procedures and Distance Measures in Data Processing

by

Robert Yuan-Shih Li

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

Thesis Director's Signature:

Shahako

Houston, Texas

May 1977
ACKNOWLEDGEMENTS

The author gratefully acknowledges the constant encouragement and generous contribution of time and energy from his advisor, Professor P. Papantoni-Kazakos. Many helpful discussions with D. Kazakos have also contributed to this thesis. The author also treasures the openness and warmth that the Kazokoses have shown toward all their students.

Helpful suggestions have also come from discussions with Professor Tapia, Professor de Figueiredo, and the author's fellow students, H. Martinez, and G. Merchant. Much love and many thanks to Aida, the author's wife, for typing this thesis.

Finally, the author wishes to acknowledge the financial support from the National Science Foundation grant ENG 75-10516 and the Air Force grant AFOSR 77-3156.
### TABLE OF CONTENTS

Abstract
Acknowledgements

**Preface**

**Part I: Optimal Sequential Procedure for Costly Data Clusters**

1. Introduction  
2. Preliminaries  
3. Bayesian Sequential Procedure  
   3.1 Two Data Clusters  
   3.2 The Two Cluster Gaussian Model  
   3.3 A Numerical Example  
4. Sequential Procedure for Unknown Priors  
   4.1 Two Data Cluster Analysis  
   4.2 The Gaussian, Two Cluster Model  
5. Absolutely Optimal Sequential Procedure for Two Data Clusters  
   5.1 Bayesian Model  
   5.2 Unknown Priors  
   5.3 Gaussian Channel Statistics  
6. Nonparametric Sequential Procedure  
   6.1 Two Cluster Case  
   6.2 Sign, Two Cluster Sequential Nonparametric Test
7. Conclusions and Suggestions for Further Research

Part II: Distance Measures and Feature Selection

1. introduction

2. Preliminaries

3. Linear Transformations

3.1 Equal Covariance Case: \( R_1 = R_2 = R \)

3.2 Equal Mean Case: \( m_1 = m_2 = m \)

4. Bhattacharry Distance; Unequal Mean and Covariance

4.1 Symmetric Functions

4.2 An Algorithm

4.3 A Numerical Example

5. Sequentially Received Data

5.1 Notations and Definitions

5.2 Equal Covariance Case: \( R_1(i) = R_2(i) = R(i) \)

5.3 Equal Mean Case: \( M(i) = 0 \)

6. Gauss-Markov Model

6.1 Linear Transformation

7. Stationary Gaussian Processes

8. Conclusions and Suggestions for Further Research

References

Appendix 1

Appendix 2

Appendix 3
PREFACE

Data processing is an integral part of many systems. These systems could be, for example, biological, social, or electrical in nature. Depending on the characteristics of the systems, various data processing techniques or procedures may be used. Obviously, in many biological systems, such as the auditory systems, the data processing procedures are complex and for the most part unknown. In this thesis, the word data means an element in some finite or infinite dimensional vector space. With this in mind, we will consider two specific data processing procedures. Specifically, in statistical communications terminology, we will be concerned with the sequential procedure; and, in pattern recognition terminology, the feature selection procedure. We emphasize here that these two procedures have wide applications and need not be restricted to problems in communications or in pattern recognition.

Since most of the data obtained in the real world contain noise as a result of observation errors or some other random variations, we will consider the data, \( x \), to be generated by a random variable. For notational sim-
plicity, we will not distinguish the random variable from the samples of the random variable. Thus, the distribution function of a random variable $x$, will be denoted by $F(x)$, and its density function by $f(x)$. Very often in statistical communications or in pattern recognition problems, one is interested in testing hypotheses concerning the data. For example, we could have two hypotheses $H_0$ and $H_1$, in which $H_0$ indicates that $f(x)$ is Gaussian with a zero mean, and $H_1$ indicates that $f(x)$ is Gaussian with a nonzero mean. The testing of a hypothesis usually involves comparing a test functional against a threshold. A hypothesis is then accepted or rejected depending on the outcome of the comparison. Both the test functional and the threshold are obtainable through the optimization of some criterion function. An often used criterion function is the probability of error, which for the two hypotheses case is a linear combination of the probabilities that the hypothesis $H_0$ is rejected when in fact $H_0$ is true, and that $H_1$ is rejected when in fact $H_1$ is true. The two coefficients for the linear combination are the a priori probabilities, or the priors, of the hypotheses $H_0$ and $H_1$ respectively.

The sequential procedures are hypotheses testing
procedures, as described above, with the additional flexibility of being able to decide that no decision on any hypothesis is possible. When this occurs, the sequential procedure then requires more data to be collected. Therefore, this procedure is well suited to hypotheses testing situations in which the data are collected sequentially. The advantage of this procedure over the procedure in which all the data have been collected, is that the sequential procedure takes fewer data points to reach a decision and therefore is faster. This becomes even more important when data collection is costly as usually happens in many practical problems. Another advantage is that in many situations it may not be easy to determine beforehand how many data points need to be collected in order for a decision to be made. Using the sequential procedure, then, the undesirable effects that the amount of data collected is insufficient or too much are circumvented. The price paid for these advantages is that sequential procedures are more complex than non-sequential procedures. However, in many situations the price paid may be more than compensated for by the advantages gained.

The concept of hypotheses testing has also been applied to pattern recognition. Since in pattern recognition the main objective is to classify patterns, it is
more natural to refer to hypotheses as classes, and to hypotheses testing as classification. The sequential procedures have been applied to pattern recognition problems as a classification method and also as a feature selection method. Feature selection generally has two distinct meanings. The first of which means the selection of a few data or features that can best represent all the data. This is similar to the concept of data reduction in communications. When used in this capacity the above method is termed intraset feature selection. The other meaning or objective of feature selection is the choosing of parts or, again, features in the data that accentuate the differences between patterns in different classes. A similar concept in communications is that of reducing the dimensionality of the data with a minimal loss of relevant information in the data. In pattern recognition terminology, this is called the interset feature selection. In this thesis, feature selection refers to interset feature selection.

The advantages of data dimensionality reduction are mainly that of reducing the computational complexity in classification. Thus, feature selection acts as a data preprocessor whose outputs are utilized by the main processor (the classifier) to facilitate classification
tasks. Also, in a lower dimensional space, more classifiers become computationally feasible allowing more flexibility in choosing the appropriate main processor. In many applications, the data collection is done in remote areas and the data must be transmitted through some channels for processing. When the available or the affordable channel capacity introduces intolerable delays in obtaining the pertinent data, then feature selection can be used at the collection site to extract the relevant information for transmission. There are also situations in which the distribution of the data needs to be estimated. By estimating the distributions or densities in a lower dimensional space, the number of training samples can be reduced. The price paid for these advantages is the loss of some information content in the original data.

This thesis will be organized in two parts. Part I will deal with the optimal sequential procedure, and Part II with the feature selection methods. The two parts will be treated independently, although as indicated, the sequential procedure can be viewed as a classification procedure; or equivalently, the feature selection procedure can be viewed as a data compression scheme.
Part I

OPTIMAL SEQUENTIAL PROCEDURE FOR COSTLY DATA CLUSTERS

1. Introduction

The sequential procedures generally can be classified as belonging to either one of two types. The first type is the sequential probability ratio test (SPRT), first formulated by Wald [1] in 1947. The second type is the Bayesian sequential procedure [2]. If we define the likelihood ratio for two simple hypotheses by

\[ T(X) = \frac{f(X/H_1)}{f(X/H_0)} \]  

(1.1)

where \( X \) is the accumulated data, then the SPRT compares \( T(X) \) against two thresholds \( A \) and \( B \) with \( A < B \). If the test is such that \( A < T(X) < B \), the procedure continues and collects new data. If \( T(X) \geq B \) or \( T(X) \leq A \), then \( H_1 \) or \( H_0 \) is decided respectively and the procedure stops. The thresholds \( A \) and \( B \) are given by

\[ A = \frac{e_{01}}{(1 - e_{10})} \]  

(1.2)

\[ B = \frac{(1 - e_{01})}{e_{10}} \]  

(1.3)

where \( e_{ij} \), with \( i, j = 0,1 \), is the error probability.
that $H_i$ is accepted, given that $H_j$ is the true hypothesis. In the Bayesian case, the sequential procedure consists of a set of decision and stopping rules. The decision rules are used to make choices among the hypotheses. The stopping rules are used to test whether the procedure should terminate at a particular step. This procedure is Bayesian because the decision rules and the stopping rules are derived by minimizing the Bayes risk.

Since the SPRT first appeared in 1947, various extensions, generalizations, alterations, and applications have been presented by many investigators [2 - 32]. Sobel [3] and Armitage [4] extended the SPRT to more than two hypotheses. Reed [5] proposed an SPRT for multi-hypotheses cases. Weiss [6] and Kiefer and Weiss [7] discussed the properties of the generalized SPRT in which the thresholds $A$ and $B$ were sequences of real numbers. Wald [1], Cox [8,9], and Armitage [10] have proposed methods to extend the SPRT to composite hypotheses by using weight functions and/or by transforming the original sequence of data into a new sequence, eliminating the nuisance parameters. Bellman et. al. [11] presented a dynamic programming procedure for the finite Bayes sequential procedure. The authors in [12 - 18] formulated and studied the SPRT in nonparametric Rank sequential tests. In applications, Dvoretzky [19], Dynkin [20],
Phataford [21], and Bhat [22] considered the sequential procedures for Markov data. Chien [23], Fu [24-26], and Patrick [27] have applied the sequential methods to problems in pattern recognition. There have also been numerous applications in communications, radar, and even social systems.

In most of the sequential procedures, the costly nature of the data is incorporated indirectly. That is, an effort is made towards the use of as few observations as possible, but no specific costs are assigned to the data [28]. Costs have been assigned to observations in some Bayesian sequential procedures [2, 29, 30]. Also, the simplicities of the SPRT are the result of choosing the start and stop boundaries A and B as in (1.2) and (1.3). The SPRT with these thresholds terminate asymptotically with Probability 1 [31,32]. Thus, the simplicity in the use of (1.2) and (1.3) is counterbalanced by the asymptotic optimal properties of the SPRT.

Recognizing that data collection costs are very real in many practical problems [33, 34], and that the thresholds in (1.2) and (1.3) may be poor for small numbers of data groups, we present an optimal sequential test for finite groups based on the SPRT. The analysis will be done for several statistical models. Emphasis will be placed on the two cluster cases. In Section 2,
the general sequential procedure is defined along with the assumptions. In Section 3, the Bayesian model is formulated. In Section 4, the case of unknown priors is considered. In Section 5, a digression to the Bayesian sequential procedure is presented for the two cluster cases. This section provides an introduction to the Bayes optimal sequential procedure, and serves to illustrate the differences and similarities between it and the procedures in Sections 3 and 4. In Section 6, the nonparametric statistical case is formulated and analyzed. In Section 7, some observations and remarks are made concerning the optimal procedure and its further research. This will conclude part I of this thesis.
2. Preliminaries

Consider a hypotheses testing procedure in which the hypotheses $H_0$ and $H_1$ are to be tested from data that are taken sequentially and in groups. The testing procedure is allowed to indicate either there is a decision in favor of one of the hypotheses or there is none. If a decision can not be made at the $i$ th step, then a new data vector $x_i$ with dimensionality $n_i$ is collected for use in the $i+1$ th step. If a decision can be made at the $i$ th step, then the procedure terminates and no more data are collected if and only if a decision for or against a hypothesis can not be made in the previous $i-1$ steps.

To use all the information contained in the data at the $i$ th step, the hypotheses testing procedure is based on evaluating the cumulative data vector $X_i$ defined below:

$$X_i = [x_{i1}^T, x_{i2}^T, \ldots, x_{in_i}^T]^T$$

(2.1)

where the superscript $T$ denotes the transpose operation, and each of the $x_j$'s is the newly obtained data vector at the $j$ th step. In this work we will assume that the maximum number of data clusters that may be collected is a finite positive integer $N$, and that to each cluster $x_j$, there is a collection cost $c_j$.

In our decision scheme we will also include penalty costs for wrong decisions. Specifically, $C_{01}$ will be the penalty paid if $H_0$ is decided while $H_1$ is true, and $C_{10}$
will be the cost for deciding on \( H_1 \) given that \( H_0 \) actually occurs. We note that the costs \( c_j, C_{01}, \) and \( C_{10} \) are all to be positive quantities.

We now define the sequential test to consist of a test functional \( T(\cdot) \), that maps the cumulative data into the reals, and three real numbers \( A, B, K \), with \( A < B \), called thresholds. The sequential scheme is then described by the relationships between the test functional and the three thresholds as follows:

i) The cumulative data \( X_i \) is formed if and only if no decision for or against a hypothesis can be made before the data cluster \( x_i \) is collected.

ii) For \( i \leq N-1 \) if the data are such that \( A < T(X_i) < B \), then no decision is possible and a new data cluster is collected.

iii) For \( i \leq N-1 \) if \( T(X_i) < A \), \( H_0 \) is decided and the procedure ends.

iv) For \( i \leq N-1 \) if \( T(X_i) > B \), \( H_1 \) is decided and the procedure ends.

v) For \( i \leq N-1 \) if \( T(X_i) = A \) or \( T(X_i) = B \) the procedure ends by deciding that \( H_0 \) is true with probability \( P_A \) or equivalently that \( H_1 \) is true with probability \( P_B \).

vi) For \( i = N \) the threshold \( K \) is used to terminate the procedure. Thus if \( T(X_N) \geq K \), \( H_1 \) is decided
and if \( T(X_N) < K \), \( H_0 \) is decided. If \( T(X_N) = K \), then \( H_1 \) is decided with probability \( P_K \).

We note here that if the distribution of the functional \( T(\cdot) \) is continuous, the probabilities \( P_A \), \( P_B \), and \( P_K \) in (v) and (vi) must be either equal to one or to zero. That is, in this case the randomization of the test is meaningless. In general, the function \( T(\cdot) \), the thresholds \( A, B, K \), and the randomization probabilities \( P_A, P_B, P_K \) are determined by the statistical information available on the data clusters. Specifically, under varying statistical knowledge about the system, the sequential procedures, considered as a tuple, \( \{ T(\cdot), A, B, K, P_A, P_B, P_K \} \), is determined through the minimization of various risk criteria.

A statistical criterion, which will be common to all of the schemes that we will analyze in this work, is the conditional average risk conditioned on the hypothesis \( H_0 \). This risk \( R_0 \) includes the penalty costs as well as the collection costs and is expressed by:

\[
R_0 = C_{10} P \left\{ \text{decide on } H_1 \text{ at some stage } / H_0 \text{ is true} \right\} \\
+ \sum_{i=1}^{N} C_i P \left\{ \text{the cluster } x_i \text{ is collected } / H_0 \text{ is true} \right\}
\]

(2.2)

where \( P[E] \) denotes the probability of the event \( E \). If we denote the event \( A < T(X_j) < B \) by \( E_j \), then from (2.2)
and from the description of the sequential test, we can obtain the following analytical expression for $R_0$:

$$R_0 = C_{10} \sum_{i=0}^{N-2} P \left\{ \bigcap_{j=1}^{i} E_j \cap \left[ \frac{T(X_{i+1}) \geq B}{H_0} \right] \right\} +$$

$$C_{10} \sum_{j=1}^{N-1} P \left\{ \bigcap_{j=1}^{N-1} E_j \cap \left[ \frac{T(X_N) \geq K}{H_0} \right] \right\} +$$

$$\sum_{i=1}^{N} C_i \sum_{j=0}^{i-1} P \left\{ \bigcap_{j=0}^{i-1} E_j / H_0 \right\}$$

(2.3)

where $\cap$ denotes intersection and $P[E_0] = 1$. We note again that if $T(X)$ has continuous distribution, the inequalities in (2.3) are to be strict.

In the following sections optimality criteria will be defined for different degrees of statistical information. The calculations of the test function and the thresholds will then be the result of minimizing these criteria. We will consider the Bayesian sequential procedure, the unknown prior sequential procedure, and the non-parametric sequential procedure. Emphasis will be placed on the two cluster situations, and illustrative examples will be given.
3. Bayesian Sequential Procedure

The Bayesian procedure assumes the availability of the complete statistical knowledge about the system. Specifically, the channel probabilities \( P_0(X) \) and \( P_1(X) \) and the priors \( q_0 \) and \( q_1 \) of the hypotheses \( H_0 \) and \( H_1 \) are known. It is natural then to choose the average risk \( R \), as the optimality criterion for the sequential test. The risk \( R \) is given by the following expression:

\[
R = q_0 R_0 + q_1 R_1
\]

where \( R_0 \) is the conditional risk expressed by (2.3) and \( R_1 \) is the risk conditioned on \( H_1 \) and given by:

\[
R_1 = C_{01} \sum_{1=0}^{N-2} P \left\{ \bigcap_{j=0}^{i} E_j \cap \left[ T(X_{i+1}) \leq A \right] \bigg/ H_1 \right\} +
\]

\[
C_{01} \sum_{j=1}^{N-1} P \left\{ \bigcap_{j=1}^{N-1} E_j \cap \left[ T(X_N) \leq K \right] \bigg/ H_1 \right\} +
\]

\[
\sum_{i=1}^{N} C_i P \left\{ \bigcap_{j=0}^{i-1} E_j \bigg/ H_1 \right\}
\]

From the expressions (2.3), (3.1), and (3.2), it is evident that the only part in \( R \) that involves the threshold \( K \) is given by:
\[ f(K) = q_0 C_{10} \sum_{j=0}^{N-1} P[\bigcap_{j=0}^{N-1} E_j] \cap \left[ T(X_N) \geq K \right| H_0] + \]
\[ q_1 C_{01} \sum_{j=0}^{N-1} P[\bigcap_{j=0}^{N-1} E_j] \cap \left[ T(X_N) \leq K \right| H_1] \quad (3.3) \]

The minimization of (3.3) will yield the threshold \( K \) and the test function \( T(X) \). By using the theorem of total probability on (3.3) we obtain:

\[ f(K) = q_0 C_{10} \int dX_N P[\text{no decision before } X_N \text{ is collected } / X_N] r(X_N) P_0(X_N) + \]
\[ q_1 C_{01} \int dX_N P[\text{no decision before } X_N \text{ is collected } / X_N] \left[1 - r(X_N)\right] P_1(X_N) \quad (3.4) \]

where the integration is over the whole space of the possible values of the cumulative data vector \( X_N \), and \( r(X_N) \) is the probability of making a decision to choose \( H_1 \) given that \( X_N \) are observed. We can prove that the optimal decision rule \( r_0(X_N) \), that minimizes the expression in (3.4), is given by:

\[
 r_0(X_N) = \begin{cases} 
 1, & X_N : \frac{P_1(X_N)}{P_0(X_N)} > \frac{q_0 C_{10}}{q_1 C_{01}} \\
 0, & X_N : \frac{P_1(X_N)}{P_0(X_N)} < \frac{q_0 C_{10}}{q_1 C_{01}} 
\end{cases} \quad (3.5)
\]
The randomization of the test does not influence the average \( f(K) \); therefore \( P_K \) can be chosen to be either one or zero. If \( P_K \) is chosen to be one, then the optimal decision function becomes:

\[
R_0(X_N) = \begin{cases} 
1, & x_N : \frac{P_1(X_N)}{P_0(X_N)} \geq \frac{q_0 c_{10}}{q_1 c_{01}} \\
0, & x_N : \frac{P_1(X_N)}{P_0(X_N)} < \frac{q_0 c_{10}}{q_1 c_{01}}
\end{cases} \tag{3.6}
\]

From (3.6) we see that for the Bayesian sequential procedure, the test and the threshold \( K \), as well as \( P_K \), can be expressed by:

\[
T(X) = \frac{P_1(X)}{P_0(X)} \tag{3.7}
\]

\[
K = \frac{q_0 c_{10}}{q_1 c_{01}} \tag{3.8}
\]

\[
P_K = 1 \tag{3.9}
\]

We can see that (3.7) is the likelihood ratio test.

The termination step at the \( N \)th data cluster is essentially the same as that of the nonsequential Bayes test with fixed data sample \( X_N \).
The thresholds \( A \) and \( B \) are determined through the minimization of the average risk \( R \) in (3.1), with the values in (3.7 - 3.9) substituted at the appropriate equations in (2.3) and (3.2). The actual calculation of \( A \) and \( B \) is plausible under certain cases and for specific channel statistics. The simplest case is when there are only two data clusters. The analysis of this case is presented below.

3.1 Two Data Clusters

Let us suppose that \( N = 2 \) in the general Bayesian expressions (2.3) and (3.2) and that \( P_K = 1 \). The expressions (3.1) for the average risk then becomes:

\[
R = q_0 c_{10} P\{T(X_1) \geq B / H_0\} + q_1 c_{01} P\{T(X_1) \leq A / H_1\} \\
+ q_0 c_{10} P\{E_1 \cap [T(X_2) \geq K] / H_0\} + \\
q_1 c_{01} P\{E_1 \cap [T(X_2) < K] / H_1\} + c_1 + \\
c_2 [q_0 P\{E_1 / H_0\} + q_1 P\{E_1 / H_1\}] 
\tag{3.10}
\]

where \( T(X) \) and \( K \) are described by (3.7) and (3.8) respectively.

For further analysis of (3.10), let us assume that
the random variables $T(X_1)$ and $T(X_2)$ are jointly absolutely continuous. Under this assumption, then, a joint density function is defined such that:

$$f_{2,1/H}(y,x)\,dy\,dx = P\{y < T(X_2) \leq y + dy, \quad x < T(X_1) \leq x + dx / H\} \quad (3.11)$$

where $H$ indicates a hypothesis. Similarly, the marginal and the conditional densities are defined by:

$$f_{1/H}(x)\,dx + P\{x < T(X_1) \leq x + dx / H\} \quad (3.12)$$

$$f_{2,1,H}(y/x) = \frac{f_{2,1/H}(y,x)}{f_{1/H}(x)} \quad (3.13)$$

respectively.

Hence, when the function $T(x)$ and the channel statistics are such that $T(X_1)$ and $T(X_2)$ are jointly absolutely continuous; and when the hypotheses are simple, the use of the equations $(3.11 - 3.13)$ in the average risk expression $(3.10)$ gives:
\[ R = R(A, B) = q_0 C_{10} \int_K \int_A f_{2,1/\mathcal{H}_0} (y,x) \, dy \, dx \\
+ q_1 C_{01} \int_K \int_B f_{2,1/\mathcal{H}_1} (y,x) \, dy \, dx \\
+ q_0 C_{10} \int_B \int_{\mathcal{H}_0} f_1 (x) \, dx \\
+ q_1 C_{01} \int_A \int_{\mathcal{H}_1} f_1 (x) \, dx \\
+ c_1 + c_2 [q_0 \int_A f_{1/\mathcal{H}_0} (x) \, dx \\
+ q_1 \int_A f_{1/\mathcal{H}_1} (x) \, dx] \]  \hspace{1cm} (3.14)

To obtain the thresholds \( A \) and \( B \), we need to minimize the above expression. If the minimum of \( R \) exists, then the thresholds \( A \) and \( B \) that minimize \( (3.14) \) are the zeros of the first order partial derivatives of \( (3.14) \) with respect to \( A \) and \( B \). Taking the derivatives gives two decoupled functions \( g(x) \) and \( h(x) \) where

\[ g(x) = \frac{\partial R(A,x)}{\partial x} , \]  \hspace{1cm} (3.15)

and

\[ h(x) = \frac{\partial R(x,B)}{\partial x} . \]  \hspace{1cm} (3.16)

If we assume that the order of integration and differentiation can be interchanged, (this requires essentially
that the integrand is bounded and measurable and that the derivative exists and is also bounded over the integration interval, then (3.15) and (3.16) become:

\[
g(x) = q_0^{C_{10}} \int_{K}^{\infty} f_{2,1/H_0}(x,y) \, dy + q_1^{C_{01}} \int_{-\infty}^{K} f_{2,1/H_1}(x,y) \, dy
- q_0^{C_{10}} f_{1/H_0}(x) + c_2 \left[ q_0 f_{1/H_0}(x) + q_1 f_{1/H_1}(x) \right] .
\]

(3.17)

\[
h(x) = - q_0^{C_{10}} \int_{K}^{\infty} f_{2,1/H_0}(x,y) \, dy
- q_1^{C_{01}} \int_{-\infty}^{K} f_{2,1/H_1}(x,y) \, dy + q_1^{C_{01}} f_{1/H_1}(x)
- c_2 \left[ q_0 f_{1/H_0}(x) + q_1 f_{1/H_1}(x) \right] .
\]

(3.18)

If we assume, in addition, that the marginal densities \( f_{1/H_0}(x) \) and \( f_{1/H_1}(x) \) do not take on zero values over the domain of \( x \), then the search for the zeros of (3.17) and (3.18) is equivalent to the search for the zeros of the following functions:
\[
G(x) = \frac{g(x)}{f_{1/H_0}(x)} = q_0 c_{10} \int_{-\infty}^{K} f_{2/1,H_0}(y/x) \, dy \\
+ q_1 c_{01} \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} \int_{-\infty}^{K} f_{2/1,H_1}(y/x) \, dy - q_0 c_{10} \\
+ c_2 q_0 + c_2 q_1 \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} 
\] (3.19)

\[
H(x) = \frac{h(x)}{f_{1/H_1}(x)} = - q_0 c_{10} \frac{f_{1/H_0}(x)}{f_{1/H_1}(x)} \int_{-\infty}^{K} f_{2/1,H_0}(y/x) \, dy \\
- q_1 c_{01} \int_{-\infty}^{K} f_{2/1,H_1}(y/x) \, dy + q_1 c_{01} \\
- c_2 q_1 - c_2 q_0 \frac{f_{1/H_0}(x)}{f_{1/H_1}(x)} 
\] (3.20)

Let us suppose that there exists a unique solution \( x^* \) to the equation \( G(x) = 0 \), and that the derivative of \( G(x) \) at \( x^* \) is greater than zero, then the threshold \( B = x^* \). Similarly, if \( y^* \) is the unique solution to the equation \( H(y) = 0 \), and if the derivative of \( H(y) \) at \( y^* \) is greater than zero, then the threshold \( A = y^* \). The above is a consequence of the facts that the derivatives of \( R \) (\( A,B \)) with respect to \( A \) and \( B \) are decoupled, and that the signs of the second derivatives of \( R \) (\( A,B \)) at the solutions are the same as the signs of the first derivatives of \( G \) and \( H \) at the solutions.

The above supposition holds when the functions \( G(x) \)
\[ \dot{G}(x) = \frac{dG(x)}{dx} = q_0 C_{10} \int_{-\infty}^{\infty} \frac{\partial f_{2/1, H_0}(y/x)}{\partial x} \, dy \]

\[ + q_1 C_{01} \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} \int_{-\infty}^{\infty} \frac{\partial f_{2/1, H_1}(y/x)}{\partial x} \, dy \]

\[ + q_1 [C_{01} \int_{-\infty}^{\infty} f_{2/1, H_1}(y/x) \, dy + c_2] \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} \, W(x) \]

(3.21)

\[ \dot{H}(x) = \frac{dH(x)}{dx} = -q_1 C_{01} \int_{-\infty}^{\infty} \frac{\partial f_{2/1, H_1}(y/x)}{\partial x} \, dy \]

\[ - q_0 C_{10} \frac{f_{1/H_0}(x)}{f_{1/H_1}(x)} \int_{-\infty}^{\infty} \frac{\partial f_{2/1, H_0}(y/x)}{\partial x} \, dy \]

\[ + q_0 [c_2 = C_{10} \int_{-\infty}^{\infty} f_{2/1, H_0}(y/x) \, dy] \frac{f_{1/H_0}(x)}{f_{1/H_1}(x)} \, W(x) \]

(3.22)

and where

\[ W(x) = \left[ \frac{df_{1/H_1}(x)}{dx} / f_{1/H_1}(x) - \frac{df_{1/H_0}(x)}{dx} / f_{1/H_0}(x) \right] \]

(3.23)
To proceed further, let us make an additional assumption that the joint statistics of $T(X_1)$ and $T(X_2)$ are such that

$$\frac{\partial f_{2/1,H}(y/x)}{\partial x} = -D \frac{\partial f_{2/1,H}(y/x)}{\partial y}$$  \hspace{1cm} (3.24)$$

where $D$ is a positive constant independent of $H$. Substituting (3.24) in (3.21) and (3.22) gives the following expressions respectively:

$$\dot{G}(x) = q_1 f_{1/H_1}(x) W(x) \left[ c_2 + c_{01} \int_{-\infty}^{K} f_{2/1,H_1}(y/x) dy \right]$$

$$+ D q_0 c_{10} f_{2/1,H_0}(K/x)$$

$$- D q_1 c_{01} \frac{f_{1/H_1}(x)}{f_{1/H_0}(x)} f_{2/1,H_1}(K/x)$$  \hspace{1cm} (3.25)$$

$$\dot{H}(x) = q_0 f_{1/H_0}(x) W(x) \left[ c_2 + c_{10} \int_{K}^{\infty} f_{2/1,H_0}(y/x) dy \right]$$

$$+ D q_1 c_{01} f_{2/1,H_1}(K/x)$$

$$- D q_0 c_{10} \frac{f_{1/H_0}(x)}{f_{1/H_1}(x)} f_{2/1,H_0}(K/x)$$  \hspace{1cm} (3.26)$$

As we mentioned before, the hypotheses $H_0$ and $H_1$ are simple in that they are each described by a fixed value.
of some parameter. If we call this parameter \( t \), then \( H_0 \) is described by the value \( t_0 \) and \( H_1 \) by \( t_1 \). Suppose now that \( t_1 > t_0 \), and that the ratio \( \frac{df_{1/t_1}(x)}{dx} / f_{1/t_1}(x) \) is a strictly monotonically increasing function of \( t \) for any fixed \( x \). In this case, we have,

\[
W(x) = \frac{\dot{f}_{1/t_1}(x)}{f_{1/t_1}(x)} - \frac{\dot{f}_{1/t_0}(x)}{f_{1/t_0}(x)} > 0 \quad \forall \ x \quad (3.27)
\]

We note that the expression (3.27) implies that

\[
\frac{d}{dx} \left( \ln \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} \right) > 0 \quad (3.28)
\]

which implies that \( f_{1/t_1}(x) \) is also strictly monotonically increasing. This monotone property further implies that

\[
\lim_{x \to \infty} \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} \to \infty \quad (3.29)
\]

and, since the ratio of the densities is non-negative, that

\[
\lim_{x \to -\infty} \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} \to 0 \quad (3.30)
\]

Before we study conditions under which \( \dot{G}(x) \) and \( \dot{H}(x) \) are positive for all \( x \), let us check the behaviors of the
functions $G(x)$ and $H(x)$ at $x = +\infty$ and at $x = -\infty$. Using (3.29) and (3.30) in (3.19) and (3.20), we obtain:

$$G(-\infty) = q_0(c_2 - c_{10}) + \lim_{x \to -\infty} q_0 c_{10} \int_{K}^{\infty} f_{2/1,t_1} (y/x) dy$$

(3.31)

$$G(+\infty) = +\infty$$

(3.32)

$$H(-\infty) = -\infty$$

(3.33)

$$H(+\infty) = -q_1(c_2 - c_{01}) - \lim_{x \to +\infty} q_1 c_{01} \int_{K}^{\infty} f_{2/1,t_1} (y/x) dy$$

(3.34)

To ensure that $G(x)$ and $H(x)$ cross the $x$-axis, we need $G(-\infty) < 0$ and $H(+\infty) > 0$. The assumption given by (3.24) implies that

$$f_{2/1,t}(y/x) = f_{2/1,t}(y - Dx)$$

(3.35)

which says that $x$ is a location parameter of the densities $f_{2/1,t_0}(y/x)$ and $f_{2/1,t_1}(y/x)$. Hence, the terms $\lim_{x \to -\infty} \int_{K}^{\infty} f_{2/1,t_0}(y/x) dy$ and $\lim_{x \to +\infty} \int_{-\infty}^{\infty} f_{2/1,t_1}(y/x) dy$ vanish when $x$ is transformed out of the integrand and into
the integration limits. We can see that if

$$c_2 < \min \left( C_{10}, C_{01} \right) \quad (3.36)$$

then $G( -\infty ) < 0$ and $H( +\infty ) > 0$.

We now return to the derivatives of $G(x)$ and $H(x)$ to examine conditions under which $\dot{G}(x)$ and $\dot{H}(x)$ are both positive for all $x$. From the expressions (3.25) and (3.26), we see that when (3.27) is true, we have the following condition on the cost $c_2$ for $G(x)$ and $H(x)$ to be both positive for all $x$:

$$c_2 > \frac{1}{\min \limits_{x} W(x)} \max \left( \max \limits_{x} \min \limits_{0_1} \left( D f_{2/1}, t_{1} \right)(K/x) $$

$$- \left[ \frac{\dot{f}_{1/t_{1}}(x)}{f_{1/t_{1}}(x)} - \frac{\dot{f}_{1/t_{0}}(x)}{f_{1/t_{0}}(x)} \right] \int_{-\infty}^{K} f_{2/1}, t_{1}(y/x) dy \right),$$

$$\max \limits_{x} C_{10} \left( D f_{2/1}, t_{0} \right)(K/x) - \left[ \frac{\dot{f}_{1/t_{1}}(x)}{f_{1/t_{1}}(x)} - $$

$$\frac{\dot{f}_{1/t_{0}}(x)}{f_{1/t_{1}}(x)} \right] \int_{-\infty}^{K} f_{2/1}, t_{0}(y/x) dy \right) \right) \quad (3.37)$$

Combining the expressions in (3.36) and (3.37) we see that if a collection cost $c_2$ exists such that
\[
\max \left( \max_x C_{01} \left( D f_{2/1,t_1}^{(K/x)} \right) - W(x) \right) \\
\frac{\left[ \int_{-\infty}^{K} f_{2/1,t_1}^{(y/x)} \, dy \right]}{\int_{-\infty}^{K} f_{2/1,t_0}^{(y/x)} \, dy} \cdot \max_x C_{10} \left( D f_{2/1,t_0}^{(K/x)} \right) - W(x) \left[ \int_{-\infty}^{K} f_{2/1,t_0}^{(y/x)} \, dy \right] / K \]
\[
\min_x W(x) < c_2 < \min \left[ C_{10}, C_{01} \right] \quad (3.38)
\]

then the functions \( G(\cdot) \) and \( H(\cdot) \) are strictly monotonically increasing, and each possesses a unique zero. From the definitions of \( G(\cdot) \) and \( H(\cdot) \), and the foregoing analysis, we conclude that the zeros minimize the average risk \( R \) in (3.14). If we call the zeros of \( H(\cdot) \) and \( G(\cdot) \), \( A \) and \( B \) respectively, then we need to check that \( A \) is less than \( B \) for them to be the thresholds defined in our sequential scheme.

The monotone properties of \( G(\cdot) \) and \( H(\cdot) \) suggest that if a value \( x_0 \) can be found such that \( G(x_0) < 0 \) and \( H(x_0) > 0 \), then we have, necessarily, that \( A < x_0 < B \). Since we are interested in the signs of \( G(\cdot) \) and \( H(\cdot) \) at \( x_0 \), we can divide (3.19) by \( q_0 C_{10} \) and (3.20) by \( q_0 C_{10} \) to obtain,
\[
\frac{f_{1/t_0}(x)}{f_{1/t_1}(x)}
\]
\[
\frac{G(x)}{q_{0}c_{10}} = c_2 \left[ \frac{1}{c_{10}} + \frac{q_1}{q_{0}c_{10}} \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} \right] + \int_{K}^{\infty} f_{2/1, t_0}(y/x) dy
\]

\[
+ \frac{q_{1}c_{01}}{q_{0}c_{10}} \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} \int_{-\infty}^{K} f_{1/1, t_1}(y/x) dy - 1 \quad (3.39)
\]

and

\[
\frac{H(x)}{q_{0}c_{10} f_{1/t_0}(x)} = - c_2 \left[ \frac{1}{c_{10}} + \frac{q_1}{q_{0}c_{10}} \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} \right]
\]

\[
- \int_{K}^{\infty} f_{2/1, t_0}(y/x) dy - \frac{q_{1}c_{01}}{q_{0}c_{10}} \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} .
\]

\[
\int_{-\infty}^{K} f_{2/1, 5_1}(y/x) dy + \frac{q_{1}c_{01}}{q_{0}c_{10}} \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} \quad (3.40)
\]

Comparing (3.39) and (3.40), it is easy to see that the
\[\text{x}_0\] that makes \(H(x_0) > 0\) and \(G(x_0) < 0\) should be such that

\[
\min \left( 1, \frac{f_{1/t_1}(x_0)}{K f_{1/t_0}(x_0)} \right) > \int_{K}^{\infty} f_{2/1, t_0}(y/x_0) dy + \frac{1}{K} \frac{f_{1/t_1}(x_0)}{f_{1/t_0}(x_0)} .
\]

\[
\int_{-\infty}^{K} f_{2/1, t_1}(y/x_0) dy + c_2 \left[ \frac{1}{c_{10}} + \frac{1}{c_{01}K} \frac{f_{1/t_1}(x_0)}{f_{1/t_0}(x_0)} \right] \quad (3.41)
\]
where the optimal threshold $K$ given by (3.8) has been substituted.

The expression in (3.41) can be viewed as another constraint on the collection cost

$$c_2 < \left( \min \left( 1, \frac{1}{K} \frac{f_{1/t_1}(x_0)}{f_{1/t_0}(x_0)} \right) - \int_{-\infty}^{\infty} f_{2/1,t_0}(y/x_0) \, dy \right)$$

$$- \frac{1}{K} \frac{f_{1/t_1}(x_0)}{f_{1/t_0}(x_0)} \int_{-\infty}^{K} f_{2/1,t_1}(y/x_0) \, dy \right) /$$

$$\left( \frac{1}{C_{10}} + \frac{1}{C_{01}} \frac{1}{K} \frac{f_{1/t_1}(x_0)}{f_{1/t_0}(x_0)} \right) \quad (3.42)$$

If we make the reasonable assumption that $c_2 > 0$, then (3.42) implies that the $x_0$ has to be such that

$$\min \left( 1, \frac{1}{K} \frac{f_{1/t_1}(x_0)}{f_{1/t_0}(x_0)} \right) - \int_{-\infty}^{\infty} f_{2/1,t_0}(y/x_0) \, dy > 0 \quad (3.43)$$

For example, if we let $x_0$ be such that

$$\frac{f_{1/t_1}(x_0)}{f_{1/t_0}(x_0)} = K \quad (3.44)$$
then (3.43) becomes:

\[ \int_{-\infty}^{\infty} f_{2/1,t}(y/x_0) \, dy + \int_{-K}^{K} f_{2/1,t}(y/x_0) \, dy \leq 1 \quad (3.45) \]

Suppose \( t \) is a location parameter for \( f_{2/1,t}(y/x) \), then incorporating (3.35) we have

\[ f_{2/1,t}(y/x_0) = f(y - Dx_0 - e(t)) \quad (3.46) \]

where \( e(*) \) is a monotonically increasing function.

Assuming densities of the form in (3.46), with the additional property that they have zero median, that is

\[ \int_{0}^{\infty} f(z) \, dz = \int_{-\infty}^{0} f(z) \, dz = 1/2 \quad (3.47) \]

then one way to satisfy (3.45) is to require that

\[ Dx_0 + e(t_0) < K < Dx_0 + e(t_1) \quad (3.48) \]

The double inequalities in (3.48) are equivalent to

\[ \frac{1}{D} (K - e(t_1)) < x_0 < \frac{1}{D} (K - e(t_0)) \quad (3.49) \]

where, as a reminder, the parameters \( t_1 \) and \( t_0 \) are such
that \( e(t_1) > e(t_0) \).

Thus, in general, if an \( x_0 \) can be found satisfying (3.43), then for a collection cost \( c_2 \) such that

\[
\max \left[ \max_x c_{01} \left( D f_{2/1,t_1}^1(K/x) - W(x) \int_{-\infty}^{K} f_{2/1,t_1}^1(y/x) \, dy \right), \right.
\]

\[
\max_x c_{10} \left( D f_{2/1,t_0}^1(K/x) - W(x) \int_{K}^{\infty} f_{2/1,t_0}^1(y/x) \, dy \right) \bigg] / \min_x W(x) < c_2 < \min \left( c_{01}, c_{10} \right) \bigg)
\]

\[
\left( \min \left( 1, \frac{1}{K} \frac{f_{1/t_1}^1(x_0)}{f_{1/t_0}^1(x_0)} \right) - \int_{K}^{\infty} f_{2/1,t_0}^1(y/x_0) \, dy \right)
\]

\[
- \frac{1}{K} \frac{f_{1/t_1}^1(x_0)}{f_{1/t_0}^1(x_0)} \int_{-\infty}^{K} f_{2/1,t_1}^1(y/x_0) \, dy \bigg) / \left( 1/c_{01} + 1/c_{10} \cdot 1/K \cdot f_{1/t_1}^1(x_0)/f_{1/t_0}^1(x_0) \right) \bigg)
\]

(3.50)

then the average risk \( R \) as expressed in (3.14) has a unique minimum when the thresholds \( A, B \) are given by the zeros of (3.22) and (3.21) respectively, and \( K \) by (3.8). We note here that if several values of \( x_0 \) satisfying (3.43) exist, then the one that provides the looser upper bound in (3.42) should be chosen.

The Bayesian sequential procedure is presented in
this section. Conditions under which the thresholds A and B exist with A < B are derived. In the following sections similar procedures for situations in which the statistical information is less complete will be covered. Two data cluster analysis will again be considered in detail. Before we proceed further with these analyses, the Bayesian two data cluster sequential method will be presented for Gaussian channel statistics.

3.2 The Two Cluster Gaussian Model

Let the data $X_1$ be Gaussian under both hypotheses $H_0$ and $H_1$, and let the hypotheses be represented by a location parameter $t$. Specifically, let the model be given by

$$X_1 = tS_i + N_1$$  

(3.51)

where $S_i = \{ s_i, i = 1,2 \}$ represents the true signals, and $N_1 = \{ n_i, i = 1,2 \}$ is a Gaussian noise vector with zero mean and covariance $R_i$. The hypotheses are represented by

$$H_0: t = t_0, \quad H_1: t = t_1 > t_0$$  

(3.52)

To apply the sequential approach, we need the joint
statistics of the noise vectors $N_1$ and $N_2$ where $N_1$ has dimensionality $k_1$ and $N_2$ has dimensionality $k_1 + k_2$. We have then $N_1 = n_1$ and $N_2 = n_2$. Let us define the cross covariance matrix between $n_2$ and $N_1$ by

$$E \left[ \begin{array}{c} n_2 \\ N_1 \end{array} \right]^T = r_{12}$$  \hspace{1cm} (3.53)

If we denote the covariance matrices of $n_1$ by $r_1$, then

$$R_2 = \begin{bmatrix} r_1 & r_{12} \\ r_{12} & r_2 \end{bmatrix} = \begin{bmatrix} r_1 & r_{12} \\ r_{12} & r_2 \end{bmatrix}$$  \hspace{1cm} (3.54)

Applying (3.7), the test functions $T(X_1)$ and $T(X_2)$ are given by:

$$T(X_1) = \frac{P_1(X_1)}{P_0(X_1)} = \exp \left( (t_1 - t_0) S_1 R_1 X_1 \right)$$

$$- \frac{1}{2} \left( t_1 - t_0 \right) S_1^T R_1^{-1} S_1$$  \hspace{1cm} (3.55)

$$T(X_2) = \frac{P_1(X_2)}{P_0(X_2)} = \exp \left( (t_1 - t_0) S_2 R_2 X_2 \right)$$

$$- \frac{1}{2} \left( t_1^2 - t_0^2 \right) S_2^T R_2^{-1} S_2$$  \hspace{1cm} (3.56)

In our sequential procedure we would test $T(X_1)$ to see if
it falls in between the thresholds A and B. If 
A < T(X_1) < B, then T(X_2) will be tested against the 
threshold K = q_0 C_{10}/q_1 C_{01}. To simplify the test func-
tions, a new threshold M will be defined by

\[ M = \frac{\ln(K)}{t_1 - t_0} + \frac{1}{2} (t_1 + t_0) S^T_{2} R_2^{-1} S_2 \]  \quad (3.57)

Thus, our new test functions become

\[ T(X_1) = S^T_1 R_1^{-1} X_1 \]  \quad (3.58)

\[ T(X_2) = S^T_2 R_2^{-1} X_2 \]  \quad (3.59)

and T(X_2) will then be used to test against M instead

of K.

The marginal density function can now be expressed by

\[ f_{1\mid t}(x) = \frac{1}{(2\pi)^{1/2} S_1 R_1 S_1^{\top}} \exp\left(-\frac{1}{2} \left( x - t S_1 R_1^{-1} S_1 \right)^T \right) \frac{1}{(2\pi)^{1/2} S_1 R_1 S_1^{\top}} \exp\left(-\frac{1}{2} \left( x - t S_1 R_1^{-1} S_1 \right)^T \right) \]  \quad (3.60)

The joint density function of T(X_1) and T(X_0) is
\[ f_{2,1/t}(y,x) = \frac{1}{(2\pi)^{\frac{1}{2}}} \exp \left( -\frac{1}{2} \left[ (y-t) S_2 R_2 S_2^T \right]^{-1} \right) \]
\[ \left[ (y-t) S_2 R_2 S_2^T, (x-t) S_1 R_1^{-1} S_1 \right]^T \]

(3.61)

where \( P \) is the 2 x 2 matrix

\[ P = \begin{bmatrix}
S_2^T R_2^{-1} S_2 & S_2^T R_2^{-1} (r_{12} R_1^{-1} S_1) \\
S_2^T R_2^{-1} (r_{12} R_1^{-1} S_1) & S_1 R_1^{-1} S_1
\end{bmatrix} \]

(3.62)

and \( |P| \) denotes the determinant of \( P \).

It is easy to show that

\[ S_1 R_1^{-1} S_1 = S_2^T R_2^{-1} (r_{12} R_1^{-1} S_1) \]

(3.63)

Using (3.13) and (3.63) the conditional density function becomes:

\[ f_{2/1,t}(y/x) = \frac{1}{(2\pi C)^{\frac{1}{2}}} \exp \left( -\frac{1}{2} \frac{(y-x-tC)^2}{C} \right) \]

(3.64)

where

\[ C = S_2^T R_2^{-1} S_2 - S_1^T R_1^{-1} S_1 \]

(3.65)
From (3.64) we easily have that

\[
\frac{\partial f_{2/1,t}^3(y/x)}{\partial x} = - \frac{\partial f_{2/1,t}^3(y/x)}{\partial y}
\]  
(3.66)

Also, it is easy to show that

\[
\frac{\dot{f}_{1/t_1}(x)}{\dot{f}_{1/t_1}(x)} - \frac{\dot{f}_{1/t_0}(x)}{\dot{f}_{1/t_0}(x)} = t_1 - t_0
\]  
(3.67)

Thus using expressions (3.57 - 3.67) in the cost \( c_2 \) expression (3.50), we obtain:

\[
\max_x \left[ \max_x \left( c_{01} \left( f_{2/1,t_1}^3(M/x) - (t_1 - t_0) \int_{-\infty}^M f_{2/1,t_1}^3(y/x) dy \right), \right. \right.
\]

\[
\left. \left. \max_x \left( c_{10} \left( f_{2/1,t_0}^3(M/x) - (t_1 - t_0) \int_{M}^{\infty} f_{2/1,t_1}^3(y/x) dy \right) \right. \right) \right]
\]

/ \( (t_1 - t_0) < c_2 < \min \left( \min (c_{01}, c_{10}) \right), \)

\[
\left( 1 - \int_{M}^{\infty} f_{2/1,t_0}^3(y/x_0) dy - \int_{-\infty}^{K} f_{2/1,t_1}^3(y/x_0) dy \right)
\]

/ \( \left( \frac{1}{c_{01}} + \frac{1}{c_{10}} \right) \)

(3.68)

in which
\[ x_0 = \frac{\ln K}{t_1 - t_0} + \frac{1}{2} (t_1 + t_0) S_T^T R_1^{-1} S_1 \quad (3.69) \]

has been picked.

Furthermore, we can show that

\[
\max_x (f_{2/1,t_1} (M/x) - (t_1 - t_0) \int_{-\infty}^{M_0} f_{2/1,t_1} (y/x) dy) =
\]

\[
\max_x (f_{2/1,t_1} (M/x) - (t_1 - t_0) \int_{M_0}^{\infty} f_{2/1,t_1} (y/x) dy) =
\]

\[
\frac{1}{(2\pi c)^{\frac{1}{2}}} \exp \left( -\frac{1}{2c^2} \right) - \frac{t_1 - t_0}{2} \text{ERFC} \left( \frac{c}{2} \right) \quad (3.70)
\]

where

\[ \text{ERFC}(Z) = 2 (\pi)^{-\frac{1}{2}} \int_{Z}^{\infty} \exp (-x^2) dx \quad (3.71) \]

The equations for the thresholds A and B are given by setting \( G(B) = 0 \) and \( H(A) = 0 \):

\[ -c_2 \left[ q_1 + q_0 \exp \left( \frac{t_0 - t_1}{2} \right) \left[ 2A - (t_1 + t_0) S_T^T R_1^{-1} S_1 \right] \right] + \theta_{01} q_1 \left[ 1 - \int_{-\infty}^{\infty} f_{2/1,t_1} (y/A) dy \right] \]

\[ - q_0 c_{10} \exp \left( \frac{t_0 - t_1}{2} \right) \left[ 2A - (t_1 + t_0) S_T^T R_1^{-1} S_1 \right] \cdot \int_{K}^{\infty} f_{2/1,t_0} (y/A) dy \quad (3.72) \]
\[ c_2 \left[ q_0 + q_1 \exp\left( \frac{t_1 - t_0}{2} \left[ 2B - (t_1 + t_0) S_1^T R_1^{-1} S_1 \right] \right) \right] \]

\[ - q_0 c_{10} \left[ 1 - \int_{M}^{\infty} f_{2/1, t_0} (y/B) dy \right] \]

\[ + q_1 c_{01} \exp \left( \frac{t_1 - t_0}{2} \left[ 2B - (t_1 + t_0) S_1^T R_1^{-1} S_1 \right] \right) \cdot \int_{-\infty}^{M} f_{2/1, t_1} (y/B) dy \]  

(3.73)

3.3 A Numerical Example

Newton's method is used to iteratively solve the equations (3.72) and (3.73) for Toeplitz covariance matrices. A matrix is said to be Toeplitz if the 
(i, j) th element of the matrix is equal to \( Q_{|i-j|} \). 
Thus, if the matrix is of order \( N \), then there can be at 
most \( N \) different values of \( Q_k \), \( k=0,1,2,..,N-1 \). For 
our example, the dimensions \( k_1 \) and \( k_2 \) are both set equal 
to 5. The covariance matrix \( R_2 \) in (3.54) is Toeplitz 
with the elements \( Q_0 = 1, Q_1 = \exp (-1), Q_2 = \exp (-2), \) 
and \( Q_3 = Q_4 = 0 \). For simplicity the parameters \( t_1 \) is 
set equal to 1, and \( t_0 \) is set equal to zero. Also, the 
time signals \( s_1 \) and \( s_2 \) are set to be ones.

The integrals \( \int_{-\infty}^{M} f_{2/1, t_0} (y/x) dy \) and \( \int_{M}^{\infty} f_{2/1, t_1} (y/x) dy \) 
are expressed through the error function \( \text{ERFC} \) by:
\[ \int_{e^{-1/1, e^{-1/3}}}^{e^{-1/1, e^{-1/3}}} (y/x) dy = \frac{1}{2} \text{ERFC} \left( \frac{e^{-1/1, e^{-1/3}}}{2c^{1/2}} \right) \] (3.74)

\[ \int_{e^{-1/1, e^{-1/3}}}^{e^{-1/1, e^{-1/3}}} (y/x) dy = \frac{1}{2} \text{ERFC} \left( \frac{e^{-1/1, e^{-1/3}}}{2c^{1/2}} \right) \] (3.75)

where \( M \) is expressed in (3.57). The expressions for \( \dot{G}(B) \) and \( \ddot{A}(A) \) in (3.21) and (3.22), coupled with the above expressions (3.74) and (3.75), are used to compute at the \( i+1 \) th iteration the values of \( A \) and \( B \) as follows:

\[ A_{i+1} = A_i - \frac{\dot{A}(A_i)}{\ddot{A}(A_i)} \] (3.76)

and

\[ B_{i+1} = B_i - \frac{\dot{G}(B_i)}{\ddot{G}(B_i)} \] (3.77)

In our computation, the cost \( C_{01} \) is fixed, and \( C_{10} \) is varied until it reaches the value of \( C_{01} \). From the expressions in (3.68) and (3.70), it is evident that the lower bound on \( c_2 \) will remain constant, independent of the ratio of \( C_{10}/C_{01} \). The upper bound of course will change when the ratio of \( C_{10} \) to \( C_{01} \) is changed. Figure 1 shows that the upper bound on \( c_2 \) increases nonlinearly with increasing values of the ratio of \( C_{10}/C_{01} \). It is worth pointing out that the bounds on \( c_2 \) as expressed in (3.68) are independent of the values of the priors.

The computed thresholds \( A \) and \( B \) are shown in Figure 2.
In general, for fixed values of the ratios $q_0/q_1$ and $C_{10}/C_{01}$, the interval size $B - A$ decreases as $c_2$ increases. The nonlinear decrease of the interval size appears as if the values of $A$ and $B$ are converging toward a common value. The interpretation for this is that as $c_2$ becomes very large, it is best to make a decision based on the first cluster. Hence, the common value would be the Bayes likelihood ratio threshold for the first cluster. For any fixed ratio of $q_0/q_1$, increasing the values of the ratio $C_{10}/C_{01}$ has the effect of increasing this common value. The interval size, however, for a fixed value $c_2$ does not seem to change as the ratio $C_{10}/C_{01}$ is varied. Increasing the values of the ratio $q_0/q_1$ has the effect of shifting the common values upward by some constant amount without altering the interval sizes. This is shown in Figure 3.

In summary, the upper bound on $c_2$ increases only with increasing values of $C_{10}/C_{01}$. The interval size seems to be the same for any fixed value of allowable $c_2$. Increasing the value of $c_2$ has the effect of decreasing the interval size; and changing the ratios of $q_0/q_1$ has the effect of shifting the common value point.
\[
c_{2}/c_{01} = 0.57 (c_{10}/c_{01})/(1 + c_{10}/c_{01})
\]

Figure 1: The bounds on \( c_2 \) as a function of \( c_{10}/c_{01} \).
Figure 2: The thresholds A, B as functions of $c_2$. 

$q_0=0.1; \frac{c_10}{c_{01}}=1$
4. **Sequential Procedure For Unknown Priors**

The statistical model here consists of known channel statistics $P_\#(X)$ and $P_0(X)$, but unknown priors $q_0$ and $q_1$. The penalty costs $C_{01}$ and $C_{01}$ and the collection costs $c_1$ are available in the present case. Consequently the risks $R_0$ and $R_1$ as defined in (2.3) and (3.2) respectively, can be computed. It is to the best interest of the designer of the receiver that both the risks remain low. An effort to minimize both costs independently will in general lead to two different sets of equations and consequently to two different sets of the thresholds $A$, $B$, and $K$. Therefore, the optimality criterion for the choice of $A$, $B$, and $K$ will be chosen to be the minimization of a risk under certain constraints.

**Definition 4.1**

Given the hypotheses $H_0$ and $H_1$, each characterized by one point $t_0$ and $t_1$ in the parameter space, the optimal sequential procedure is defined as the one that minimizes the conditional cost $R_1$ under the constraint that the conditional cost $R_0$ remains below a given constant $C_0$.

In our development of the sequential procedure, we
will require only that $R_0 = C_0$ to avoid making the test unnecessarily tight. If the hypothesis $H_0$ is characterized by a set of values $T_0$ in the parameter space $T$, then the following version of the optimal sequential test definition will be used.

**Definition 4.2**

Given the hypotheses $H_0$ and $H_1$, where $H_0$ is characterized by the parameter space $T_0$, and $H_1$ consists of one point $t_1$ in the parameter space, then the optimal sequential procedure is defined as the one that minimizes $R_1$ under the constraint that $\sup_{T_0} R_0 = C_0$.

**Definition 4.3**

If the optimal sequential procedure obtained from choosing a fixed value $t_0 \in T_0$ remains the same for any choice of $t \in T_0$, then this procedure is called uniformly optimal within $T_0$.

We will now consider the sequential problem when the hypotheses are simple as described in definition 4.1. The constraint on $R_0$ suggests the use of Lagrange multiplier method. Specifically, the optimal sequential procedure $(T(X), A, B, K)$ is the one that minimizes
\[ R(\lambda) = R_1 - \lambda R_0 \]  \hspace{1cm} (4.1)

under the condition that \( A < B \). In (4.1), \( \lambda \) is the Lagrange multiplier which is to be determined, after (4.1) has been minimized, from

\[ R_0(A(\lambda), B(\lambda), K(\lambda)) = C_0 \] \hspace{1cm} (4.2)

In minimizing (4.1), the inequality constraint \( A < B \) will be treated as if it is not there. It is hoped that through the strict minimization of (4.1), \( A \) will be less than \( B \). Conditions similar to those in Section 3 will be developed to ensure that \( A < B \).

As in Section 3, the threshold \( K \) and the test function will be derived first. From the expressions (2.3), (3.1), and (3.2), the part of the risk \( R \) that involves \( K \) is given by:

\[
f(K, \lambda) = C_{01} p \left\{ \left[ \bigcap_{j=0}^{N-1} E_j \right] \cap \left[ T(X_N) \leq K \right] \big/ H_1 \right\} - \lambda C_{10} p \left\{ \left[ \bigcap_{j=0}^{N-1} E_j \right] \cap \left[ T(X_N) \geq K \right] \big/ H_0 \right\} \] \hspace{1cm} (4.3)

The function \( f(K, \lambda) \) in (4.3) can also be expressed by:
\[ f(K,\lambda) = C_{01} P \{ \text{deciding on } H_0 \text{ when the } N \text{ th data cluster is obtained } / H_1 \} \]

\[ -\lambda C_{10} P \{ \text{deciding on } H_1 \text{ when the } N \text{ th data cluster is obtained } / H_0 \} \quad (4.4) \]

Applying the theorem of total probability on (4.4) using the data \( X_N \) as the variable, we obtain

\[ f(K,\lambda) = C_{01} \int dX_N P \{ \text{no decision before } X_N \text{ is collected} / X_N \} \left[ 1 - r(X_N) \right] P_1(X_N) \]

\[ -\lambda C_{10} \int dX_N P \{ \text{no decision before } X_N \text{ is collected} / X_N \} r(X_N) P_0(X_N) \quad (4.5) \]

where the integration ranges over the values that the random variable \( X_N \) may take. The function \( r(X_N) \) is the probability that hypothesis \( H_1 \) is decided on given that the cumulative data \( X_N \) are available. This is similar to the Bayesian case, and as before the optimal decision rule \( r_0(X_N) \) is:

\[
r_0(X_N) = \begin{cases} 
1 & \frac{P_1(X_N)}{P_0(X_N)} \geq \frac{-\lambda C_{10}}{C_{01}} \\
0 & \frac{P_1(X_N)}{P_0(X_N)} < \frac{-\lambda C_{10}}{C_{01}} 
\end{cases} \quad (4.6)
\]
It is easily seen from (4.6) that

$$T(X) = \frac{P_1(X)}{P_0(X)}$$  \hspace{1cm} (4.7)$$

$$K = \frac{-\lambda C_{10}}{C_{01}}$$  \hspace{1cm} (4.8)$$

where is to be determined from (4.2).

Again, as in Section 3, unless a specific statistical model is given, it is not possible to find $A(\lambda)$ and $B(\lambda)$ for an arbitrary number of clusters $N$. However, if $N = 2$, then analysis under fairly general assumptions can be done. We present below the two cluster analysis.

### 4.1 Two Data Cluster Analysis

Let $N = 2$, then the expression for $R(\lambda)$ in (4.1) becomes:

$$R(\lambda) = C_{01} P \{ E_1 \cap [T(X_2) \leq K] / H_1 \}$$

$$+ \lambda C_{10} P \{ E_1 \cap [T(X_2) \geq K] / H_0 \} + \phi_1 (1-\lambda)$$

$$+ C_2 \left[ -\lambda P[E_1 / H_0] + P[E_1 / H_1] \right] C_{01} P[T(X_1]$$

$$\leq A / H_1 \right) - \lambda C_{10} P \{ T(X_1) \geq B / H_0 \}$$  \hspace{1cm} (4.9)$$

where the optimal $T(X)$ and $K$ are given by (4.7) and (4.8).
respectively. For further analysis, we will assume as in Section 3.1 that the random variables \( T(X_1) \) and \( T(X_2) \) are jointly absolutely continuous. The joint, marginal, and condition density functions are as defined in (3.11 - 3.13) respectively. Substituting these density functions and the optimal \( K \) into (4.9) gives:

\[
R(\lambda, A, B) = c_1 \int_{t_0}^{t_1} \lambda \, \mathcal{E}_0 \int_{-\infty}^{K} \int_{A}^{B} f_{2,1/t_0}(y,x) \, dy \, dx +
\int_{-\infty}^{K} \int_{A}^{B} f_{2,1/t_1}(y,x) \, dy \, dx +
\int_{-\infty}^{A} \int_{-\infty}^{B} f_{1,t_0}(x) \, dx + \int_{A}^{B} f_{1,t_1}(x) \, dx \]

\[
(4.10)
\]

The optimal thresholds \( A \) and \( B \), when they exist, are to be found by setting the gradients of (4.10), with respect to \( A \) and \( B \), to zero. This is because we are searching for \( A \) and \( B \) over the open set \( (-\infty, \infty) \). The derivatives of (4.10), with respect to \( A \) and \( B \), yield two decoupled functions \( h(\cdot) \) and \( g(\cdot) \) where
\[ H(\lambda, x) = \frac{\partial R(\lambda, x, B)}{\partial x} / f_1/t_1(x) = -c_2 + c_2 \lambda \frac{f_1/t_0(x)}{f_1/t_1(x)} + 
\]
\[ + c_{01} - c_{01} \int_{-\infty}^{K} f_{2/1, t_1}(y/x) dy + \lambda c_{10} \frac{f_1/t_0(x)}{f_1/t_1(x)} \int_{-\infty}^{K} f_{2/1, t_0}(y/x) dy \quad (4.11) \]

and

\[ G(\lambda, x) = \frac{\partial R(\lambda, A, x)}{\partial x} / f_1/t_0(x) = -c_2 \lambda + c_2 \frac{f_1/t_1(x)}{f_1/t_0(x)} + \]
\[ + \lambda c_{10} - \lambda c_{10} \int_{K}^{\infty} f_{2/1, t_0}(y/x) dy + c_{01} \frac{f_1/t_1(x)}{f_1/t_0(x)} \int_{-\infty}^{K} f_{2/1, t_1}(y/x) dy \quad (4.12) \]

The threshold \( K \) is given by (4.8) in the above expressions (4.11) and (4.12).

As in Section 3, if the functions \( G(\lambda, x) \) and \( H(\lambda, x) \) are for some \( \lambda \)'s strictly monotonically increasing with \( x \) from a negative to a positive value, then they both have one zero each. The zeros are \( A(\lambda) \) and \( B(\lambda) \) that minimize (4.10) if \( A(\lambda) < B(\lambda) \). Let us now suppose that \( t_1 > t_0 \), and that the ratio \( f_1/t(x) / f_1/t(x) \) is a strictly monotonically increasing function. The last condition implies that (3.28) and (3.29) hold. If we make a further
assumption that \( x \) is a location parameter so that
\[
f_{2/1, t}(x) = f_{2/1, t}(y - Dx) \quad \text{where D is a positive constant, then we have:}
\]
\[
G(\lambda, -\infty) = -\lambda(c_2 - C_{10}) \quad (4.13)
\]
\[
G(\lambda, +\infty) = +\infty \quad (4.14)
\]
\[
H(\lambda, -\infty) = \text{SGN}(\lambda) \cdot \infty \quad (4.15)
\]
\[
H(\lambda, +\infty) = C_{01} - c_2 \quad (4.16)
\]
The expressions (4.13 - 4.16) imply that for \( G(\lambda, x) \) and \( H(\lambda, x) \) to be as desired, the Lagrange multiplier must be such that
\[
\lambda < 0 \quad (4.17)
\]
and that the collection cost \( c_2 \) should be
\[
c_2 < \min \{ C_{01}, C_{10} \} \quad (4.18)
\]
With (4.17) and (4.18) in mind, let us now investigate the behaviors of the derivatives of \( G(\lambda, x) \) and \( H(\lambda, x) \) to ascertain conditions under which the thresholds are the minimizers of the risk in (4.10). We assume as before that the order of integration and differentiation can be interchanged. If we let \( \dot{G}(\lambda, x) = \frac{\partial G(\lambda, x)}{\partial x} \) and
\[ H(\lambda, x) = \frac{\partial H(\lambda, x)}{\partial x}, \text{ then we have:} \]

\[ G(\lambda, x) = \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} W(x) \left[ c_2 + C_{01} \int_{-\infty}^{K} f_{2/1, t_1}(y/x) dy \right] 
- C_{01} D \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} f_{2/1, t_1}(K/x) 
+ \lambda C_{10} D f_{2/1, t_0}(K/x) \quad (4.19) \]

\[ H(\lambda, x) = -\lambda \frac{f_{1/t_0}(x)}{f_{1/t_1}(x)} W(x) \left[ c_2 + C_{10} \int_{K}^{\infty} f_{2/1, t_0}(y/x) dy \right] 
+ \lambda C_{10} \frac{f_{1/t_0}(x)}{f_{1/t_1}(x)} D f_{2/1, t_0}(K/x) 
+ C_{01} D f_{2/1, t_1}(K/x) \quad (4.20) \]

where \( W(x) \) is defined by (3.23) and (3.27).

Analogous to the Bayesian analysis, we have that if the costs \( C_{10}, C_{01}, \) and \( c_2 \) are such that

\[ \max_x \left( \max_x C_{10} (D f_{2/1, t_0}(K/x) - W(x) \int_{K}^{\infty} f_{2/1, t_1}(y/x) dy) \right), \]

\[ \max_x C_{01} (D f_{2/1, t_1}(K/x) - W(x) \int_{-\infty}^{\infty} f_{2/1, t_1}(y/x) dy) / \]

\[ \min_x W(x) \geq c_2 < \min (C_{01}, C_{10}) \quad (4.21) \]
then \(G(\lambda,x)\) and \(H(\lambda,x)\) will be positive for all \(x\) and all \(\lambda < 0\). What (4.17) and (4.21) imply is that the function \(R(\lambda,A,B)\) achieves its minimum at \(A(\lambda)\) and \(B(\lambda)\), which are the zeros of (4.20) and (4.19) respectively. It remains to study the relationship between the two constants \(A(\lambda)\) and \(B(\lambda)\).

Since the functions \(H(\lambda,x)\) and \(G(\lambda,x)\) are monotone under the conditions of (4.17) and (4.18), the existence of a value \(x_0(\lambda)\), such that \(G(\lambda,x_0(\lambda)) < 0\) and \(H(\lambda,x_0(\lambda)) > 0\), implies that \(A(\lambda) < x_0(\lambda) < B(\lambda)\) which satisfies the constraint in (4.1). From (4.11) and (4.12) we obtain for \(\lambda < 0\)

\[
\frac{G(\lambda,x)}{-C_{10}^{\lambda}} = -c_2 \left[ \frac{1}{C_{10}^{\lambda}} + \frac{1}{-C_{10}^{\lambda}} \frac{f_{1/t_0}(x)}{f_{1/t_1}(x)} \right] + \int_{-\infty}^{K} f_{2/1,t_0}(y/x) dy - 1 \quad (4.22)
\]

\[
\frac{H(\lambda,x)}{-C_{10}^{\lambda}} = -c_2 \left[ \frac{1}{C_{10}^{\lambda}} + \frac{1}{-C_{10}^{\lambda}} \frac{f_{1/t_0}(x)}{f_{1/t_1}(x)} \right] - \int_{K}^{\infty} f_{2/1,t_0}(y/x) dy + \frac{C_{01}}{C_{10}^{\lambda}} \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} \quad .
\]

\[
\int_{-\infty}^{K} f_{2/1,t_1}(y/x) dy - \frac{C_{01}}{C_{10}^{\lambda}} \frac{f_{1/t_1}(x)}{f_{1/t_0}(x)} \quad (4.23)
\]
Expressions (4.22) and (4.23) and (4.17) imply that if an \( x_0(\lambda) \) exists, then it should necessarily satisfy the following inequality if the condition \( c_2 > 0 \) is taken into account:

\[
\min \left( 1, \frac{1}{K(\lambda)} \frac{f_{1/t_1}(x_0(\lambda))}{f_{1/t_0}(x_0(\lambda))} \right) > \int_{K(\lambda)}^{\infty} f_{2/1, t_0}(y/x_0(\lambda)) \, dy \\
+ \frac{1}{K(\lambda)} \frac{f_{1/t_1}(x_0(\lambda))}{f_{1/t_0}(x_0(\lambda))} K(\lambda) \int_{-\infty}^{f_{2/1, t_0}(y/x_0(\lambda))} \, dy
\]

(4.24)

the condition on \( c_2 \) then becomes:

\[
c_2 < \min \left( 1, \frac{1}{K(\lambda)} \frac{f_{1/t_1}(x_0(\lambda))}{f_{1/t_0}(x_0(\lambda))} \right) - \int_{K(\lambda)}^{\infty} f_{2/1, t_0}(y/x_0(\lambda)) \, dy \\
- \frac{1}{K(\lambda)} \frac{f_{1/t_1}(x_0(\lambda))}{f_{1/t_0}(x_0(\lambda))} K(\lambda) \int_{-\infty}^{f_{2/1, t_0}(y/x_0(\lambda))} \, dy \right) / \\
\left( \frac{1}{c_{10}} + \frac{1}{c_{01} K(\lambda)} \frac{f_{1/t_1}(x_0(\lambda))}{f_{1/t_0}(x_0(\lambda))} \right)
\]

(4.25)

Combining (4.21) and (4.25) we again have that if
\[
\max \left[ \max_x C_{10} (D f_{2/1, t_0} (K(\lambda)/x) - W(x) \int_{K(\lambda)}^{\infty} f_{2/1, t_0} (y/x) dy) \right] \\
\max_x C_{01} (D f_{2/1, t_1} (K(\lambda)/x) - W(x) \int_{-\infty}^{\infty} f_{2/1, t_1} (y/x) dy) \right] / \\
\min_x W(x) < c_2 < \min [C_{01}, C_{10}],
\]

\[
\left( \min \left( 1, \frac{f_{1/t_1} (x_0(\lambda))}{K(\lambda) \int_{K(\lambda)}^{\infty} f_{2/1, t_0} (y/x_0(\lambda)) dy} \right) - \frac{f_{1/t_1} (x_0(\lambda))}{K(\lambda) \int_{K(\lambda)}^{\infty} f_{2/1, t_0} (y/x_0(\lambda)) dy} \right) / \\
\left( \frac{1}{C_{10}} + \frac{1}{C_{01} K_0(\lambda) \int_{K(\lambda)}^{\infty} f_{1/t_1} (x_0(\lambda)) dx} \right) \right] \tag{4.26}
\]

for some \( x_0 \) satisfying (4.24), then the function \( R(\lambda, A, B) \) has the unique minimizer \( A(\lambda) \) and \( B(\lambda) \) which when substituted into the expression for \( R_0 \) should, hopefully, yield a \( \lambda \) that is negative. The expression for \( R_0(\lambda, A(\lambda), B(\lambda)) \) is

\[
R_0(\lambda) A(\lambda), B(\lambda)) \ast C_{10} \int_{B(\lambda)}^{\infty} f_{1/t_0} (x) dx + \\
C_{10} \int_{A(\lambda)}^{B(\lambda)} \int_{K(\lambda)}^{\infty} f_{2,1/t_0} (y/x) dy dx + c_1 \\
+c_2 \int_{A(\lambda)}^{B(\lambda)} f_{1/t_0} (x) dx = c_0 \tag{4.27}
\]
The above procedure will be better understood by an example involving specific statistics. Following the example, we will digress to discuss the absolutely optimal sequential procedure before we present the nonparametric sequential procedure.

4.2 The Gaussian Two Cluster Model

We use the same model presented in Section 3.2 except for the fact that the priors \( q_0 \) and \( q_\Phi \) are unknown. The analysis in Section 3.2 carries over here except that the threshold \( K \) is now defined by (4.8) with a negative valued \( \lambda \). Also the equations in (3.72) and (3.73) are replaced by:

\[
c_2 \left[ -\lambda + \exp \left( \frac{t_0 - t_0}{2} (2B(\lambda) - (t_1 + t_0) S_1^T R_1^{-1} S_1) \right) \right] \\
+ 2\pi \epsilon_{10} \left[ 1 - \int_{K(\lambda)}^{\infty} f_{2/1,t_0}(y/B(\lambda))dy \right] \\
+ c_{01} \exp \left( \frac{t_1 - t_0}{2} (2B(\lambda) - (t_1 + t_0) S_1^T R_1^{-1} S_1) \right) \\
\int_{-\infty}^{K(\lambda)} f_{2/1,t_1}(y/B(\lambda))dy = 0 \quad (4.28)
\]
\[-c_2 \left[ 1 - \lambda \exp \left( \frac{t_0 - t_1}{2} (2A(\lambda) - (t_1 - t_0) S_1^T R_1^{-1} S_1) \right) \right] \]

\[+ c_{01} \left[ 1 - \frac{K(\lambda)}{\int_{-\infty}^{t_1} f_{2/1} (y/A(\lambda)) \, dy} \right] \]

\[+ \lambda c_{10} \exp \left( \frac{t_0 - t_1}{2} (2A(\lambda) - (t_1 + t_0) S_1^T R_1^{-1} S_1) \right) \cdot \]

\[K(\lambda) \int_{f_{2/1}, t_0} (y/A(\lambda)) \, dy = 0 \quad (4.29) \]

where \( \lambda \) is the solution of \((4.27)\).
5. Absolutely Optimal Sequential Procedure For Two Data Clusters

In this section we will examine the derivation of a sequential procedure without defining the procedure a priori as we did in Section 2. The procedure will be developed for the two cluster case where the data are still described by vectors $X_1$ as defined in (2.4). We retain also the collection costs $\theta_i$ and the penalty costs $C_{10}$ and $C_{01}$. The hypotheses $H_0$ and $H_1$ are assumed to be simple, each consisting of a single parameter value $t_0$ and $t_1$ respectively. Our first step will be to express the conditional costs $R_0$ and $R_1$ in a similar fashion as in Section 2. We obtain, therefore, that:

$$R_0 = C_{10} P \{ \text{decide on } H_1 / H_0 \} + c_1 P \{ \text{stop at cluster } 1 / H_0 \} + (c_1 + c_2) P \{ \text{stop at cluster } 2 / H_0 \} \tag{5.1}$$

$$R_1 = C_{01} P \{ \text{decide on } H_0 / H_1 \} + c_1 P \{ \text{stop at cluster } 1 / H_0 \} + (c_1 + c_2) P \{ \text{stop at cluster } 2 / H_1 \} \tag{5.2}$$

Introducing the data vectors $X_1$ and $X_2$ as random
variables in the above expressions and defining:

\[ D_i(X_i) = P \{ \text{stop at cluster } i \text{ / observed } X_i \} \quad (5.3) \]

= stopping rule

\[ d_i(X_i) = P \{ \text{decide on } H_i \text{ / stopped at cluster } i, \text{ observed } X_i \} \quad (5.4) \]

= decision rule

we obtain:

\[
R_0 = C_{10} \int_1^d x_1 d_1(x_1) D_1(x_1) P_0(x_1) + \\
C_{10} \int_2^d x_2 d_2(x_2) D_2(x_2) P_0(x_2) + \\
c_1 + c_2 \int_2^d D_2(x_2) P_0(x_2) dx_2 \quad (5.5)
\]

\[
R_1 = C_{01} \int_1^d x_1 [1 - d_1(x_1)] D_1(x_1) P_1(x_1) + \\
C_{01} \int_2^d x_2 [1 - d_2(x_2)] D_2(x_2) P_1(x_2) + \\
c_1 + c_2 \int_2^d D_2(x_2) P_1(x_2) dx_2 \quad (5.6)
\]

where the integration regions 1 and 2 are the ranges of the random variables \(X_1\) and \(X_2\) respectively.

The expressions (5.1) and (5.2) imply a sequential procedure that first decides a stopping time and then
decides on whether to make a decision on one of the hypotheses. For the two cluster case, if a decision on one of the hypotheses can not be made, then \( x_2 \) will have to be collected to form \( \mathbf{x}_2 = [X_1^T, X_2^T]^T \). Mathematically this means that

\[
D_2(X_2) = 1 - D_1(X_1) \quad (5.7)
\]

The equations in (5.5) and (5.6) now become:

\[
R_0 = C_{10} \int dX_1 d_1(X_1) D_1(X_1) P_0(X_1) + \\
C_{10} \int dX_1 [1 - D_1(X_1)] \int dX_2 d_2(X_2) P_0(X_2) + \\
c_1 + c_2 \int dX_1 [1 - D_1(X_1)] P_0(X_1) \quad (5.8)
\]

\[
R_1 = C_{01} \int dX_1 [1 - d_1(X_1)] D_1(X_1) P_1(X_1) + \\
C_{01} \int dX_1 [1 - D_1(X_1)] \int dX_2 [1 - d_2(X_2)] P_1(X_2) + \\
c_1 + c_2 \int dX_1 [1 - D_1(X_1)] P_1(X_1) \quad (5.9)
\]

where the integration region \( \mathcal{J} \) refers to the range of the random variable \( x_2 \).

The optimal choice of the stopping rule \( D_1(X_1) \) and the decision rules \( d_1(X_1) \) and \( d_2(X_2) \) varies with the statistical model. In the following sections, the Bayesian model and the unknown priors case will be analyzed.
5.1 **Bayesian Model**

In this model the stopping rule and the decision rules are to be chosen through the minimization of the average risk,

\[ R = q_0 R_0 + q_1 R_1 \]  \hspace{1cm} (5.10)

where \( q_0 \) and \( q_1 \) are the priors of the hypotheses \( H_0 \) and \( H_1 \) respectively. Substituting (5.8) and (5.9) in (5.10) we get:

\[
R = \int_1 dX_1 d_1(X_1) D_1(X_1) \left[ q_0 C_{10} P_0(X_1) - q_1 C_{01} P_1(X_1) \right] \\
+ q_1 C_{01} + c_1 + c_2 + \int_1 dX_1 \left[ 1 - D_1(X_1) \right] \\
\int_3 dx_2 d_2(x_2) \left[ q_0 C_{10} P_0(x_2) - q_1 C_{01} P_1(x_2) \right] \\
- \int_\infty dx_1 D_1(X_1) c_2 \left[ q_0 P_0(X_1) + q_1 P_1(X_1) \right] \]  \hspace{1cm} (5.11)

Observing that the decision rules \( d_1(X_1) \) and \( d_2(x_2) \) are applied at different stages of the procedure, they are independent of each other. Therefore we will seek the optimal decision rules first. From (5.11) we see that \( d_1(X_1) \) appears only in the first term. Denoting this term by \( f_1(d_1(X_1)) \) we have:
\[ f_1(d_1(x_1)) = \int_1^{x_1} d_1(x_1) d_1(x_1) \left[ q_0c_{10}p_0(x_1) - q_1c_{01}p_1(x_1) \right] \]  

(5.12)

The optimal decision rule \( d_1^0(x_1) \) is expressed in the following theorem.

**Theorem 5.1**

The rule \( d_1^0(x_1) \) expressed by

\[
d_1^0(x_1) = \begin{cases} 
1 ; x_1 : & \frac{P_1(x_1)}{P_0(x_1)} > \frac{q_0c_{10}}{q_1c_{01}} \\
P_1 ; x_1 : & \frac{P_1(x_1)}{P_0(x_1)} = \frac{q_0c_{10}}{q_1c_{01}} \\
0 ; x_1 : & \frac{P_1(x_1)}{P_0(x_1)} < \frac{q_0c_{10}}{q_1c_{01}} 
\end{cases}
\]

(5.13)

minimizes (5.12) over any other rule \( d_1(x_1) \). Here \( P_1 \) is a positive number \( \in [0,1] \).

**Proof:**

Let \( d_1(x_1) \) be any other rule then

\[
f_1(d_1(x_1)) - f_1(d_1^0(x_1)) = \int_1^{x_1} d_1(x_1) \left[ d_1(x_1) - d_1^0(x_1) \right] 
\]

\[
D_1(x_1) \left[ q_0c_{10}p_1(x_1) \right]
\]
\[
\begin{align*}
= \int_{x_1} \left[ 1 - a_1(x_1) \right] d_1(x_1) \left[ q_1 c_{01} p_1(x_1) - q_0 c_{10} p_0(x_1) \right] \\
&\quad + \int_{x_1} d_1(x_1) d_1(x_1) \left[ q_0 c_{10} p_0(x_1) - q_1 c_{01} p_1(x_1) \right] \\
&\quad + \int_{x_1} \left[ d_1(x_1) - p_1 \right] d_1(x_1) \left[ q_0 c_{10} p_0(x_1) - q_1 c_{01} p_1(x_1) \right] \geq 0
\end{align*}
\]

Similarly the decision rule \( d_2(x_2) \) appears in (5.11) in the term

\[
f_2(d_2(x_2)) = \int \frac{d_2(x_2)}{2} d(x_2) \cdot
\]

\[
\left[ q_0 c_{10} p_0(x_2) - q_1 c_{01} p_1(x_2) \right] 
\]

and we can express the optimal decision rule \( d_2^0(x_2) \) by the following theorem.

**Theorem 5.2**

The rule \( d_2^0(x_2) \) expressed by
\[
\begin{align*}
\mathcal{D}_2^O(X_2) &= \begin{cases}
1 &; x_2 : \frac{p_1(x_2)}{p_0(x_2)} > \frac{q_0 c_{10}}{q_1 c_{01}} \\
0, 1 &; x_2 : \frac{p_1(x_2)}{p_0(x_2)} = \frac{q_0 c_{10}}{q_1 c_{01}} \\
0 &; x_2 : \frac{p_1(x_2)}{p_0(x_2)} < \frac{q_0 c_{10}}{q_1 c_{01}}
\end{cases}
\end{align*}
\]

minimizes (5.14) over any other rule \( d_2(x_2) \).

**Proof**

The same as in theorem 5.1.

With the optimal decision rules specified, we are now in a position to study the stopping rule \( D_1(X_1) \). The terms in (5.11) involving \( D_1(X_1) \) are gathered to form

\[
P_3(D_1(X_1)) = \int_1^x dx_1 D_1(x_1) \left[ d_1(x_1)[q_0 c_{10} p_0(x_1) - q_1 c_{01} p_1(x_1)] \right]
- \int_0^x dx_2 d_2(x_2)[q_0 c_{10} p_0(x_2) - q_1 c_{01} p_1(x_2)]
- c_2 [q_0 p_0(x_1) + q_1 p_1(x_1)]
\]  

(5.16)

The optimal stopping rule \( D_1^O(X_1) \) will be the one that minimizes (5.16). Following the same logic as in theorem 5.1, we can state that the optimal stopping rule can be expressed by:
\[ D^O_1(X_1) = \begin{cases} 
1 ; & X_1 : \int_3^x \text{dx}2 \text{d}2(x_2) \left[ q_0C_{10}P_0(X_2) - q_1C_{01}P_1(X_2) \right] + c_2 \left[ q_0P_0(X_1) + q_1P_1(X_1) \right] \\
> d_1(X_1) \left[ q_0C_{10}P_0(X_1) - q_1C_{01}P_1(X_1) \right] \\
Q_1 \in [0,1] ; & X_1 : \text{holds in the expression} \\
0 ; & X_1 : < \text{holds in the expression (5.17)} 
\end{cases} \]

By substituting the optimal decision rules into (5.17) we can expand \( D^O_1(X_1) \) into:

For \( X_1 : \frac{P_1(X_1)}{P_0(X_1)} > \frac{q_0C_{10}}{q_1C_{01}} \)

\[ D^O_1(X_1) = \begin{cases} 
1 ; & \frac{P_1(X_1)}{P_0(X_0)} > q_0 \left[ C_{10} \left[ 1 - \int_2^x \text{dx}2 \text{d}2P_0(x_2/X_1) \right] - c_2 \right] \\
& q_1 \left[ C_{01} \left[ 1 - \int_2^x \text{dx}2P_1(x_2/X_1) \right] + c_2 \right] \\
& \equiv \text{expression 1} \\
\end{cases} \]

Region of integration is for \( x_2 \) such that

\[ \frac{P_1(X_2)}{P_0(X_2)} > \frac{q_0C_{10}}{q_1C_{01}} \]

For \( Q_1 \in [0,1] \)

\[ \frac{P_1(X_0)}{P_0(X_0)} = \text{expression 1} \]

\[ 0 ; \frac{P_1(X_0)}{P_0(X_0)} < \text{expression 1} \]
For $X_1$: 

$$ \frac{P_1(X_1)}{P_0(X_1)} = \frac{\frac{q_0 c_{10}}{q_1 c_{01}}}{1 + \frac{q_0 [c_2 - c_{01} \int dx_2 P_1(x_2/X_1)]}{q_0 [c_2 + c_{10} \int dx_2 P_0(x_2/X_1)]}$$

$= \text{expression 2}$

Region of integration is for $x_2$ such that

$$P_1^0(X_1) = \begin{cases} 
\frac{P_1(X_2)}{P_0(X_2)} > \frac{q_0 c_{10}}{q_1 c_{01}} & \text{expression 2} \\
Q_1 \in [0, 1] & \frac{P_1(X_0)}{P_1(X_1)} \\
0 & \frac{P_1(X_0)}{P_1(X_1)} < \text{expression 2}
\end{cases}$$

(5.18)

where $P_1$ and $P_2$ have been chosen to be zero.

We observe that the optimal stopping rule in (5.18) can be simplified further only for specific channel statistics. Before we illustrate the above procedure for Gaussian statistics, we will consider the unknown priors case.

5.2 **Unknown Priors**

In the unknown priors case the optimality criterion
will be the minimization of the risk \( R_1 \) under the restriction that \( R_0 \) equals some fixed value \( C_0 \). Again, we use the Lagrange multiplier method by introducing the multiplier \( \lambda \) to minimize the risk \( R(\lambda) \) defined by

\[
R(\lambda) = R_1 - \lambda R_0 \tag{5.19}
\]

where \( \lambda \) is to be determined from

\[
R_0(\lambda) = C_0 \tag{5.20}
\]

We observe that the formulation of the problem is the same as in the Bayesian case if we make the substitutions \( q_1 = 1 \) and \( q_0 = -\lambda \). Hence we have immediately that the optimal decision rules \( d_1^o(X_1) \) and \( d_2^o(X_2) \) are:

\[
d_1^o(X_1) = \begin{cases} 
1 & X_1 : \frac{P_1(X_1)}{P_0(X_1)} > -\lambda \frac{C_{10}}{C_{01}} \\
P_1 & X_1 : = \text{ holds} \\
0 & X_1 : < \text{ holds} \tag{5.21}
\end{cases}
\]

\[
d_2^o(X_2) = \begin{cases} 
1 & X_2 : \frac{P_1(X_2)}{P_0(X_2)} > -\lambda \frac{C_{10}}{C_{01}} \\
P_2 & X_2 : = \text{ holds} \\
0 & X_2 : < \text{ holds} \tag{5.22}
\end{cases}
\]
where \( P_1 \) and \( P_2 \) are numbers in the interval \([0,1]\). The optimal stopping rule \( D_1^0(X_1) \) is also similar to (5.18) except that the ratio \( q_0 / q_1 \) is replaced by \(-\lambda\). Let us consider now the sequential procedure for Gaussian statistics.

5.3 Gaussian Channel Statistics

Let the statistics be as follows:

\[
\begin{align*}
P_0(X_1) &= G(0, R_1) \\
P_1(X_1) &= G(S_1, R_1) \\
P_0(X_2) &= G(0, R_2) \\
P_1(X_2) &= G(S_2, R_2)
\end{align*}
\]

(5.23)

where \( G(x, y) \) indicates Gaussian density with mean \( x \) and covariance \( y \). From Section 3.2, we know that the test \( \frac{P_1(X_1)}{P_0(X_1)} \) against the threshold \( q_0 C_{10} / q_1 C_{01} \) is equivalent to the test of \( S_1^T R_1^{-1} X_1 \) against \( \ln(q_0 C_{10} / q_1 C_{01}) + 1/2 S_1^T R_1^{-1} S_1 \). Similarly, the test \( \frac{P_1(X_1)}{P_0(X_2)} \) against \( q_0 C_{10} / q_1 C_{01} \) is equivalent to testing \( S_2^T R_2^{-1} X_2 \) against \( \ln(q_0 C_{10} / q_1 C_{01}) + 1/2 S_2^T R_2^{-1} S_2 \). Hence, the
integral \( \int dx_2 \, P_1(x_2/x_1) \) is the probability that
\[
\frac{p_1(x_2)}{p_0(x_2)} > \frac{q_{0C10}}{q_{1C01}}
\]

\( S_2^T R_2^{-1} x_2 \) is greater than \( \ln(\frac{q_{0C10}}{q_{1C01}}) + 1/2 S_2^T R_2^{-1} S_2 \)
given that \( x_1 \) has been collected and that the hypothesis
\( H_i \), \( i=0,1 \), is true.

To include both the Bayesian case and the case of
the unknown priors, let us define \( K \) to be either
\( q_{0C10}/q_{1C01} \) or \( -\lambda C_{10}^{-1} C_{01} \) corresponding to the two
cases. Hence the tests \( S_1^T R_1^{-1} x_1 \) and \( S_2^T R_2^{-1} x_2 \) will be against the thresholds \( \ln K + 1/2 S_1^T R_1^{-1} S_1^{-1} \) and
\( \ln \frac{K}{K} + 1/2 S_2^T R_2^{-1} S_2 \) respectively.

For simplicity, let
\[
R_1 = I_{k_1}, \quad R_2 = I_{k_1+k_2}
\]
\[
S_1 = \{ s_i = s > 0 ; \ i=1,\ldots,k_1 \}
\]
\[
S_2 = \{ s_i = s > 0 ; \ i=1,\ldots,k_1+k_2 \}
\]
\[
X_1 = \{ z_i ; \ i=1,\ldots,k_1 \}
\]
\[
X_2 = \{ z_i ; \ i=1,\ldots,k_1+k_2 \}
\]

where \( k_1 \) and \( k_2 \) are the data cluster dimensions.

We can now express the integral
\[ \int dx_2 \frac{p_1(x_2/x_1)}{p_0(x_2)} = p \left\{ \sum_{i=1}^{k_1+k_2} \frac{\ln(K)}{s} + \frac{s(k_1+k_2)}{2} / x_1, H_1 \right\} \]

\[
= p \left\{ \sum_{i=k_1+1}^{k_1+k_2} \ln(K) + \frac{s(k_1+k_2)}{2} - \sum_{i=1}^{k_1} \frac{1}{k_2} \sum_{i=1}^{k_1} z_i, H_i \right\} \]

\[
= \begin{cases} 
1 - \phi \left( \frac{\ln(K)}{sk_2} + \frac{s(k_1+k_2)}{2k_2} - \frac{1}{k_2} \sum_{i=1}^{k_1} z_i \right), & H_i = H_0 \\
1 - \phi \left( \frac{\ln(K)}{sk_2} + \frac{s(k_1-k_2)}{2k_2} - \frac{1}{k_2} \sum_{i=1}^{k_1} z_i \right), & H_i = H_1 
\end{cases} \]

(5.29)

where

\[
\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp \left( - \frac{u^2}{2} \right) \, du \quad (5.30)
\]

For notational simplicity let

\[
h_0 = \frac{\ln(K)}{sk_2} + \frac{s(k_1+k_2)}{2k_2} - \frac{1}{k_2} \sum_{i=1}^{k_1} z_i \quad (5.31)
\]

\[
h_1 = \frac{\ln(K)}{sk_2} + \frac{s(k_1-k_2)}{2k_2} - \frac{1}{k_2} \sum_{i=1}^{k_1} z_i \quad (5.32)
\]

\[
z_0 = \frac{sk_1}{2} + \frac{\ln K}{s} + \frac{1}{s} \ln \left( \frac{\phi(h_0) - c_2/C_{10}}{\phi(h_1) + c_2/C_{01}} \right) \quad (5.33)
\]

\[
z_1 = \frac{sk_1}{2} + \frac{\ln K}{s} + \frac{1}{s} \ln \left( \frac{c_2/C_{10} + 1 - \phi(h_0)}{c_2/C_{01} - 1 - \phi(h_1)} \right) \quad (5.34)
\]

Applying (5.18) the optimal stopping rule for this example becomes:
\[ D_1^0(x_1) = \begin{cases} 
\text{For } x_1 : & k_1 \sum_{i=1}^{k_1} z_i > \frac{\ln(K)}{s} + \frac{sk_1}{2} \\
1 & \text{If } \phi(h_0) > c_2/c_{10} \\
 & k_1 \sum_{i=1}^{k_1} z_i > Z_0 \\
 & \text{If } \phi(h_0) \leq c_2/c_{10} \\
 & \text{no other test needed} \\
Q_1 \in [0,1] & \phi(h_0) > c_2/c_{10} \\
 & k_1 \sum_{i=1}^{k_1} z_i = Z_0 \\
0 & \phi(h_0) > c_2/c_{10} \\
 & k_1 \sum_{i=1}^{k_1} z_i < Z_0 \\
\text{For } x_1 : & k_1 \sum_{i=1}^{k_1} z_i \leq \frac{\ln(K)}{s} + \frac{k_1 s}{2} \\
1 & \text{If } \phi(h_1) \geq 1 - c_2/c_{01} \\
 & \text{no other test needed} \\
 & \text{If } \phi(h_1) < 1 - c_2/c_{01} \\
 & k_1 \sum_{i=1}^{k_1} z_i < Z_1 \\
Q_1 \in [0,1] & \phi(h_1) < 1 - c_2/c_{01} \\
 & k_1 \sum_{i=1}^{k_1} z_i = Z_1 \\
0 & \phi(h_1) < 1 - c_2/c_{01} \\
 & k_1 \sum_{i=1}^{k_1} z_i > Z_1 \\
\end{cases} \]

(5.35)
The above stopping rule is diagrammed in Figure 4.

It is obvious from (5.35) that for the Bayes case one can simply substitute the $z_i$'s to obtain the stopping rule. For the unknown priors case the problem can only be studied parametrically. Once the $z_i$'s are available the selection of different values of $K$ can lead to different optimal stopping rules. One then needs to compute the conditional risk $R_0$ to see if (5.20) is satisfied. We note here that even for the Bayesian case the optimal stopping rule requires more computations on the observed data than the sequential procedure developed in the previous sections.

We mention here also, that the procedure presented in this section is essentially similar to the truncated sequential procedure (Ferguson [2]). The main difference is that we consider the data to be taken in groups, and that the collection costs are assigned to the groups, not to the individual elements of the group.
Figure 4: Optimal Stopping Rule for Two Cluster Case
6. **Nonparametric Sequential Procedures**

In this section we consider statistical models that are described nonparametrically. That is, the knowledge about the channel statistics is almost nil, and is described only by the general characteristics of a nonparametric family. The costly nature of the data suggests again the use of sequential procedures which, in this case, should maintain nonparametric properties.

The characteristic that will be asked to remain unchanged inside the whole nonparametric family will be the conditional risk $R_0$ given in (2.2). Randomization parameters $P_A$ and $P_B$ defined by

$$P_A = P \text{ decide on } H_0/\bar{T}(X_1) = A \quad (6.1)$$

$$P_B = P \text{ decide on } H_1/\bar{T}(X_1) = B \quad (6.2)$$

, however, may be introduced in (2.2). With the above description, the definition of a sequential nonparametric test is given below.

**Definition 6.1**

A sequential test as defined in Section 2 is called nonparametric inside a given nonparametric family, if and only if the conditional risk $R_0$ is the same for the
whole family.

Clearly, if the sequential test consists of a function \( T(.) \) whose statistics under hypothesis \( H_0 \) are independent of the particular distribution, then it is nonparametric. The thresholds \( A, B, \) and \( K, \) being characteristics of the test, should remain the same inside the nonparametric family. Therefore, their evaluations must be based exclusively on the statistical behavior conditioned on \( H_0 \). Optimality measures of such statistical behavior could be the conditional risk \( R_0 \) as well as its lower or upper bounds. To analyze the problem in detail we will concentrate on the two cluster case, and specify conditions under which the thresholds can be determined.

6.1 Two Cluster Case

The conditional risk \( R_0 \) in (2.2), after introducing the randomization parameters \( P_A \) and \( P_B \) becomes:

\[
R_0 = C_{10} P\{ T(X_1) > B/H_0 \} + C_{10} P_B P \{ T(X_1) = B/H_0 \} +
C_{10} P \{ E_1 \cap T(X_2) > K/H_0 \} + C_{10} P_K P \{ E_1 \cap T(X_2) = K/H_0 \} +
C_{10} (1 - P_A) P \{ T(X_1) = A \cap T(X_2) > K/H_0 \} +
C_{10} (1 - P_A) P \{ T(X_1) = A \cap T(X_2) = K/H_0 \} +
\]
\[ C_{10}(1 - P_B) P \{ T(X_1) = B \cap T(X_2) = K/H_0 \} + \]
\[ C_{10}(1 - P_B) P \{ T(X_1) = B \cap T(X_2) = K/H_0 \} + c_1 + \]
\[ c_2 P \{ E_1/H_0 \} + c_2(1 - P_A) P \{ T(X_1) = A/H_0 \} + \]
\[ c_2(1 - P_B) P \{ T(X_1) = B/H_0 \} \quad (6.3) \]

The randomization parameters \( P_A, P_B, \) and \( P_K \) are necessary if some fixed values are given for the risk \( R_0 \) and for the bounds on \( R_0 \). A natural lower bound on \( R_0 \) is the risk \( R_{01} \) that the system takes if a decision is made after the data vector \( X_1 \) has been collected. This risk is given by

\[ R_{01} = c_1 + C_{10} P \{ T(X_1) > B/H_0 \} + P_B P \{ T(X_1) = B/H_0 \} \]
\[ (6.4) \]

An upper bound on \( R_0 \) can be obtained by combining the risk \( R_{01} \) and the risk \( R_{02} \), where \( R_{02} \) is the risk the system takes in making a non-sequential decision based on the fixed data vector \( X_2 \). We can express \( R_{02} \) by:

\[ R_{02} = c_1 + c_2 + C_{10} P \{ T(X_2) > K/H_0 \} + C_{10} P K P \{ T(X_2) = \]
\[ K/H_0 \} \quad (6.5) \]

Thus the bounds on \( R_0 \) are given by:

\[ R_{01} < R_0 < R_0 + R_{02} \quad (6.6) \]
If the risks $R_{01}$, $R_{02}$, and $R_0$ that can be afforded by the system are available to the designer, then the set \{ $A$, $B$, $K$, $P_A$, $P_B$, $P_K$ \} is well defined and so is the sequential nonparametric test. Although the risk $R_0$ does not carry the identification of the particular distribution in the nonparametric family that characterizes the channel, the risk $R_1$ does. Therefore, $R_1$ can be used as an evaluation measure of the nonparametric test for the channel distributions inside the nonparametric family. Similarly, as in nonsequential nonparametric tests, the calculation of the risks $R_0$ and $R_1$ are generally not feasible analytically for finite sample sizes. Therefore, an asymptotic relative evaluation measure characterizing a particular distribution will be assigned to each sequential nonparametric two cluster test. This measure will be called, as in the nonsequential case, the asymptotic relative efficiency. Formally, we define this measure below.

**Definition 6.2**

Let the data clusters be of equal size, and let the hypotheses $H_0$ and $H_1$ approach each other, then the asymptotic relative efficiency is the ratio of the cluster size $K_p$, required by the parametric test to achieve the
specified risks \( R_0 \) and \( R_1 \), over the cluster size \( K_n \) required by the nonparametric test to achieve the same risks. It is understood that the evaluation measure is to be performed at a given distribution.

The parametric sequential test that will be used as a basis for evaluation in the above definition will, in general, be the unknown priors presented in Section 4. We will calculate the asymptotic relative efficiency of several two cluster sequential nonparametric tests using the above basis. The sequential nonparametric tests will be characterized by functions \( T(.) \) that are generally used in the nonsequential nonparametric tests.

### 6.2 Sign, Two Cluster Sequential Nonparametric Test

The sequential sign test consists of a function

\[
T(X) = \frac{1}{n} \sum_{i=1}^{n} SGN(z_i)
\]

(6.7)

where

\[
X = \{ z_i, \ i=1,...,n \}
\]

(6.8)

and

\[
SGN(Z) = \begin{cases} 
1 & Z \geq 0 \\
0 & Z < 0 
\end{cases}
\]

(6.9)

and a set \( \{ A, B, K, P_A, P_B, P_K \} \) that is determined from the specified values of \( R_0, R_{01}, \) and \( R_{02} \). To find the
asymptotic relative efficiency of the sign test with respect to the unknown priors, parametric test, let the underlying distribution be Gaussian. Specifically, let the hypotheses be described by

\[ H_0 : z_i ; \quad i=1, \ldots, \quad z_i \text{ is } G(0, \sigma) \]

\[ H_1 : z_i ; \quad i=1, \ldots, \quad z_i \text{ is } G(t, \sigma) \]

where \( t > 0 \) and the data \( z_i \)'s are assumed to be independent. The condition \( H_0 \rightarrow H_1 \) in this case is equivalent to \( t \rightarrow 0 \). To achieve low risks \( R_0 \) and \( R_1 \) under the assumption that \( t \rightarrow 0 \), the cluster sizes will be allowed to increase to infinity. We let the two clusters have one common size \( K_n \), for the nonparametric test and \( K_p \) for the corresponding parametric test. Rewriting (6.7) for the nonparametric case, we have:

\[
T(X_n) = \frac{1}{K_n} \sum_{i=1}^{K_n} \text{SGN}(z_i) \quad (6.10)
\]

The statistics of the function \( T(X_n) \) under \( H_0 \) and \( H_1 \) are:

\[
T(X_n)/H_0 = G(0, \frac{1}{(K_n)^\frac{1}{2}}) \quad (6.11)
\]

\[
T(X_n)/H_1 = G(\phi(\frac{t}{\sigma}) - \phi(-\frac{t}{\sigma}), \frac{1}{K_n^{\frac{1}{2}}})
\]

\[
\xrightarrow{t \rightarrow 0} G(\frac{2\phi^{\frac{1}{2}}}{n} \frac{t}{\sigma} ; \frac{1}{K_n^{\frac{1}{2}}}) \quad (6.12)
\]
Similarly,

$$T(X_2)/H_0 = G(0, \frac{1}{(2K_n)^{\frac{1}{2}}})$$  \hspace{1cm} (6.13)

$$T(X_2)/H_1 \xrightarrow{t \to 0} G(\frac{2}{n}\frac{t}{\sigma}, \frac{1}{(2K_n)^{\frac{1}{2}}})$$  \hspace{1cm} (6.14)

We have also that $T(X_1)$ and $T(X_2)$ are jointly Gaussian, with cross covariance under $H_0$ equal to

$$E[T(X_1)T(X_0)/H_0] = \frac{1}{2K_n} \sum_{j=1}^{K_n} \sum_{i=1}^{K_n} [SGN(z_1) \cdot SGN(z_1)/H_0]$$

$$= \frac{1}{2K_n}$$  \hspace{1cm} (6.15)

From (6.3 - 6.6) and (6.11 - 6.15) it is easy to obtain:

$$R_0 = c_{10} [1 - \phi(B(K_n)^{\frac{1}{2}})] + c_{10} \int_{-K_n}^{K_n} B \int_{-\frac{B}{K_n}}^{\frac{B}{K_n}} e^x \exp \left(-K_n \cdot (2y^2 - 2xy + x^2)\right) dy + c_1 + c_2 \left[ \phi(B(K_n)^{\frac{1}{2}}) - \phi(A(K_n)^{\frac{1}{2}}) \right]$$  \hspace{1cm} (6.16)

$$R_{01} = c_1 + c_{10} \left[1 - \phi(B(K_n)^{\frac{1}{2}})\right]$$  \hspace{1cm} (6.17)

$$R_{02} = c_1 + c_2 + c_{10} \left[1 - \phi(K(2K_n)^{\frac{1}{2}})\right]$$  \hspace{1cm} (6.18)

where $\phi$ is as defined by (5.30).
If values for $R_0$, $R_{01}$ and $R_{02}$ are given, then the thresholds $A, B, K$ are specified uniquely. In a similar way, we can express the risk $R_1$ by:

$$R_1 = C_{01} \phi \left( \left[ A - \left( \frac{\hat{z}_n}{\sigma} \right)^{\frac{1}{2}} \frac{t}{\sigma} \right] (K_n)^{\frac{1}{2}} \right) +$$

$$C_{01} \int_{-\infty}^{K} \int_{A}^{B} \frac{K_n}{n} \exp \left\{ -K_n \left[ 2 \left( y - \left( \frac{\hat{z}_n}{\sigma} \right)^{\frac{1}{2}} \frac{t}{\sigma} \right)^2 \right. \right.$$

$$- 2 \left. \left( x - \left( \frac{\hat{z}_n}{\sigma} \right)^{\frac{1}{2}} \frac{t}{\sigma} \right) \left( y - \left( \frac{\hat{z}_n}{\sigma} \right)^{\frac{1}{2}} \frac{t}{\sigma} \right) \right]\right\} + c_1 +$$

$$c_2 \left\{ \phi \left( \left[ A - \left( \frac{\hat{z}_n}{\sigma} \right)^{\frac{1}{2}} \frac{t}{\sigma} \right] (K_n)^{\frac{1}{2}} \right) - \phi \left( \left[ B - \left( \frac{\hat{z}_n}{\sigma} \right)^{\frac{1}{2}} \frac{t}{\sigma} \right] (K_n)^{\frac{1}{2}} \right) \right\}$$

(6.19)

To calculate the asymptotic efficiency of the sign test relative to the Gaussian distribution, we also need to find the $R_0, R_1$ expressions for the unknown priors two cluster parametric test. The parametric test for the same hypotheses consists of a function

$$T_1(X) = \frac{1}{n} \sum_{i=1}^{n} z_i$$

(6.20)

and a set of thresholds $(A_1, B_1, K_1)$ as described in Section 4. The statistics of the function $T_1(X_1)$ and $T_1(X_2)$ under the assumption that each cluster has $K_p$ elements are:

$$- T_1(X_1)/H_0 = G(0, \frac{\sigma}{(K_p)^{\frac{1}{2}}})$$

$$T_1(X_1)/H_1 = G(t, \frac{\sigma}{(K_p)^{\frac{1}{2}}})$$

(6.21)
\[ T_1(X_2) / H_0 = G(0, \frac{\sigma}{(2K_p)^{1/2}}) \quad (6.23) \]

\[ T_1(X_2) / H_1 = G(t, \frac{\sigma}{(2K_p)^{1/2}}) \quad (6.24) \]

The cross covariance of \( T_1(X_1) \) and \( T_1(X_2) \) equals \( \frac{\sigma^2}{2K_p} \), and

\[ T_1(X_2) / T_1(X_1) = \frac{Z^* H_0}{} \overset{G(\frac{Z}{2}, \frac{\sigma}{2(K_p)^{1/2}})}{=}. \quad (6.25) \]

\[ T_1(X_2) / T_1(X_1) = \frac{Z^* H_1}{} \overset{G(\frac{Z+t}{2}, \frac{\sigma}{2(K_p)^{1/2}})}{=} \quad (6.26) \]

Applying the above expressions to (6.3) and (3.2), we obtain:

\[ R_0 = C_{10} \left[ 1 - \varnothing \left( \frac{B_1}{\sigma} (K_n)^{1/2} \right) \right] + \]

\[ C_{10} \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dx \frac{K_p}{\pi \sigma^2} \exp \left\{ - \frac{K_p}{\sigma^2} [2y^2 - 2xy + x^2] \right\} + c_1 \]

\[ + c_2 \left[ \varnothing \left( \frac{B_1}{\sigma} (K_p)^{1/2} \right) \right] - \varnothing \left( \frac{A_1}{\sigma} (K_p)^{1/2} \right) \quad (6.27) \]

\[ R_1 = C_{01} \varnothing \left( \left[ A_1 - t \right] \frac{K_p}{\sigma} \right) + c_1 + c_2 [\varnothing (\left[ B_1 - t \right] \frac{K_p}{\sigma}) - \varnothing \left( \left[ A_1 - t \right] \frac{K_p}{\sigma} \right)] - \]

\[ \varnothing (\left[ A_1 - t \right] \frac{K_p}{\sigma}) + C_{01} \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dx \frac{K_p}{\pi \sigma^2} \exp \left\{ - \frac{K_p}{\sigma^2} \left[ 2(y-t)^2 - 2(x-t)(y-t) + (x-t)^2 \right] \right\} \quad (6.28) \]
To obtain the ratio $K_p/K_n$ it is necessary to go through an iterative process in which, for a fixed value of $t$, small relative to the deviation, an initial value for $K_p$ is estimated and substituted into (6.27) and (6.28). Equations similar to those of (4.11) and (4.12) are then used in conjunction with (6.27) to obtain a set of values for $A_1, B_1, K_1$. Finally, these values are substituted into (6.28) to check if the risk $R_1$ for the parametric case is the same as the risk $R_1$ for the nonparametric case. If not, a new $K_p$ is picked, and the process repeats. A simpler way, for our example, is to do the reverse. Specifically, we first select reasonable values for $K_p$ and $t$, which will enable us to obtain the values $A_1, B_1, K_1$, and, hence, $R_1$. We then use this $R_1$ in the nonparametric case to compute the values of $A, B, K,$ and $K_n$. Before plunging into this procedure, we may want to check that the costs are such that the parametric case will be meaningful.

Recall from Section 4, that the optimal threshold $K_1$ is

$$K_1 = \left| \lambda \right| \frac{c_{10}}{c_{01}}$$

(6.29)

and that $A_1(\lambda)$ and $B_1(\lambda)$ are solutions to
\[ c_2 \left| \lambda \right| + \exp \left\{ \frac{tK_p}{2\sigma^2} \left[ 2B_1(\lambda) - t \right]\right\} - \left| \lambda \right| c_{10} \varphi \left( \frac{2(K_p)^{\frac{1}{2}}}{\sigma} \left[ K_1(\lambda) - \frac{B_1(\lambda) + t}{2} \right]\right) = 0. \]  

\[ \frac{K_1(\lambda) - \frac{B_1(\lambda) + t}{2}}{2} = 0. \]  

\[ -c_2 \left| \lambda \right| \exp \left\{ -\frac{tK_p}{2\sigma^2} \left[ 2A_1(\lambda) - t \right]\right\} + c_{10} \left[ 1 - \varphi \left( \frac{2(K_p)^{\frac{1}{2}}}{\sigma} \left[ K_1(\lambda) - \frac{A_1(\lambda) + t}{2} \right]\right) \right] = 0 \]  

(6.31)

Following the analysis in Section 4, the condition for \( A_1(\lambda) \) and \( B_1(\lambda) \) to exist is:

\[ c_2 < \min \left( c_{10}, c_{10} \right) \]  

(6.32)

The condition for \( A_1(\lambda) \) to be less than \( B_1(\lambda) \) is:

\[ c_2 < \min \left( 1, \frac{1}{K_1(\lambda)} \exp \left\{ \frac{tK_p}{2\sigma^2} \left[ 2x_0(\lambda) - t \right]\right\} \right) - 1 + \varphi\left( \frac{2(K_p)^{\frac{1}{2}}}{\sigma} \left[ K_1(\lambda) - \frac{x_0(\lambda)}{2} \right]\right) - \frac{1}{K_1(\lambda)} \exp \left\{ -\frac{tK_p}{2\sigma^2} \left[ 2x_0(\lambda) - t \right]\right\} \varphi \left( \frac{2(K_p)^{\frac{1}{2}}}{\sigma} \left[ K_1(\lambda) - (x_0(\lambda) + t)/2 \right]\right) \]  

\[ \left( \frac{1}{c_{10}} + \frac{1}{c_{10}} \frac{1}{K_1(\lambda)} \exp \left\{ \frac{tK_p}{2\sigma^2} \left[ 2x_0(\lambda) - t \right]\right\} \right) \]  

(6.33)
for some $x_0$ that makes (6.33) positive. Such an $x_0$
could be

$$x_0 = \frac{c^2}{tkp} \ln K_1(\lambda) + \frac{t}{2} \tag{6.34}$$

which when substituted in (6.33) gives:

$$c_2 < \phi \left( \frac{2(K_p)^{1/2}}{\sigma} \left[ K_1(\lambda) - \frac{2}{2tkp} \ln K_1(\lambda) - \frac{t}{4} \right] \right) -$$

$$\phi \left( \frac{2(K_p)^{1/2}}{\sigma} \left[ K_1(\lambda) - \frac{2}{2tkp} \ln K_1(\lambda) - \frac{3t}{4} \right] \right) /$$

$$\left( \frac{1}{c_{01}} + \frac{1}{c_{01}} \right) \tag{6.35}$$

To ensure that $A_1(\lambda)$ and $B_1(\lambda)$ are minimizers of the risk
$R = R_1 - \lambda R_0$, let us consider the equations derived from
(4.19) and (4.20):

$$\dot{G}(x, t) = c_2 \frac{tkp}{\sigma} \exp \left\{ \frac{tkp}{2\sigma^2} [2x-t] \right\} \left[ c_2 + c_{01} \phi \left( \frac{2(K_p)^{1/2}}{\sigma} \right\}$$

$$\left[ K_1(\lambda) - \frac{x+t}{2} \right] \right) - c_{01} \frac{(K_p)^{1/2}}{\sigma} \exp \left\{ \frac{tkp}{2\sigma^2} [2x-t] \right\}$$

$$g\left( \frac{2(K_p)^{1/2}}{\sigma} \left[ K_1(\lambda) - \frac{x+t}{2} \right] \right) + |\lambda| c_{010} \left( \frac{K_p}{\sigma} \right)^{1/2} g\left( \frac{2(K_p)^{1/2}}{\sigma} \left[ K_1(\lambda) - \frac{x}{2} \right] \right) \tag{6.36}$$
\[
\dot{H}(\lambda, x) = |\lambda| \frac{tK_p}{\sigma^2} \exp \left\{ -\frac{tK_p}{2\sigma^2} [2x-t] \right\} \left[ c_2 + c_{10} - c_{10} \phi \left( \frac{2(K_p)^{\frac{1}{2}}}{\sigma} \left[ K_1(\lambda) - \frac{x}{2} \right] \right) + c_{01} \frac{(K_p)^{\frac{1}{2}}}{\sigma} \phi \left( \frac{2(K_p)^{\frac{1}{2}}}{\sigma} \left[ K_1(\lambda) - \frac{x}{2} \right] \right) \right) - |\lambda| c_{10} \frac{(K_p)^{\frac{1}{2}}}{\sigma} \exp \left\{ -\frac{tK_p}{2\sigma^2} [2x-t] \right\} g(\frac{2(K_p)^{\frac{1}{2}}}{\sigma} \left[ K_1(\lambda) - \frac{x}{2} \right]) \right) \tag{6.37}
\]

where \( g(x) = \frac{1}{(2\pi)^{\frac{1}{2}}} \exp \left\{ -\frac{x^2}{2} \right\} \).

We see that since \( K_p \) is supposedly large, the term \( \phi (\cdot) \) approached 1, and hence, if

\[
c_2 > \frac{\sigma}{t(2\pi K_p)^{\frac{1}{2}}} \max (c_{01}, c_{10}) \tag{6.38}
\]

then both \( G(\lambda, x) \) and \( \dot{H}(\lambda, x) \) will be greater than zero.

We must observe here that the quantity \( t(K_p)^{\frac{1}{2}} \) in (6.38) must not approach zero.

Combining (6.32), (6.35), and (6.38), we then have that for

\[
\frac{\sigma \max(c_{01}, c_{10})}{t(2\pi K_p)^{\frac{1}{2}}} < \sigma_2 < \min \left( \left( \phi \left( \frac{2(K_p)^{\frac{1}{2}}}{\sigma} \left[ K_1(\lambda) - \frac{\sigma^2 \ln K_1(\lambda)}{2tK_p} \right] - \frac{t}{4} \right) \right) - \phi \left( \frac{2(K_p)^{\frac{1}{2}}}{\sigma} \left[ K_1(\lambda) - \frac{\sigma^2 \ln K_1(\lambda)}{2tK_p} \right] - \frac{3t}{4} \right) \right) / \left( \frac{1}{c_{10}} + \frac{1}{c_{01}} \right), \ c_{10}, \ c_{01} \right) \tag{6.39}
\]
where $K_1(\lambda)$ is given by (6.29), there exist unique constants $A_1(\lambda)$ and $B_1(\lambda)$ for the parametric sequential test.
7. **Conclusions and Suggestions for Further Research**

In Part I of this thesis, we have developed an optimal sequential probability ratio test that includes the collection costs of a finite number of data clusters. In Section 2, the sequential procedure was defined. In Section 3, the procedure was developed for the Bayes statistical model. In Section 4, the unknown priors case was considered. In Section 5, the optimal truncated Bayes sequential procedure was discussed to contrast it with our procedure. In Section 6, the optimal two cluster sequential probability ratio test was defined for non-parametric cases. A definition of asymptotic relative efficiency was also given.

Our development in this part of the thesis has concentrated primarily on the two cluster case. We have obtained bounds on the collection cost $c_2$ that will guarantee the existence of the thresholds $A$ and $B$. Practical examples using Gaussian statistics have been presented to illustrate the sequential procedures.

The selection of the point $x_0$ for the upper bound has been arbitrary. It will be interesting to see if the $x_0$ we selected is the best $x_0$. Since the extension
of our two cluster analysis to higher cluster numbers becomes extremely complicated, even for specific densities, we may consider a merging procedure in which the first n-1 clusters are merged into one cluster, (if no decision has been made), and the n-th cluster then becomes cluster 2. This sequential procedure appears to be practical and should be investigated.
Part II

DISTANCE MEASURES AND FEATURE SELECTION

1. Introduction

The main aim of feature selection is to reduce the dimensionality of the data while at the same time preserving as much class discriminatory information as possible. To be useful, a feature selection procedure should also be easy to implement. Mathematically, feature selection can be viewed as a problem in finding an operator from a high-dimensional space to a low dimensional space such that a performance functional is extremized.

One of the most stringent performance functionals is the probability of error. Let us consider a random variable to belong to one of two possible classes, \( H_0 \) and \( H_1 \) with priors \( q_0 \) and \( q_1 \), respectively. If we denote the conditional probability density functions of the random variable under \( H_0 \) and \( H_1 \) by \( f_0(\cdot) \) and \( f_1(\cdot) \), respectively, then the probability of error (\( P_e \)) is expressed by:

\[
P_e = \int q_0 f_0(x)dx + \int q_1 f_1(x)dx \quad \text{(1.1)}
\]

where the integration regions 0 and 1 are the decision
regions inside which $H_0$ and $H_1$, respectively, are accepted. The direct minimization of $P_e$ is often impossible, even in the presence of simple statistics [36]. de Figueiredo [35], Kazakos [36], and Guseman [37] have considered algorithms for feature selection by minimizing the probability of error. Their computational procedures, however, are complex.

The difficulties of using $P_e$ for feature selection have led many investigators [38-50] to use simpler, more tractable, criteria. The majority of these criteria fall in the category of distances of probability measures. A list of some of the more commonly used distance measures is given by Kanal [51]. The attractiveness of a distance measure generally depends on its calculability and on its power as a measure of $P_e$. Since many distances involve expressions of the probability measures, they may be sensitive to changes in the probability measures. Thus, the sensitivity of a distance should be considered when the underlying probability measures are not well known.

In this thesis we will examine several distances and their relationships to each other. One distance, the Bhattacharyya distance, will be considered in detail. Since the goal of feature selection is to find appropriate operators that are simple to implement, we will consider
only linear operators, and will represent them by linear transformations.

In Section 2, we will present and compare the Chernoff, Bhattacharyya, I-divergence, Vasershtein, Kolmogorov, Variational, Prokhorov, and Levy distances. In Section 3, the above distances will be used to derive linear transformations for two-class Gaussian distributed data. We will consider both the equal mean and the equal covariance cases. In Section 4, the Bhattacharyya distance will be used to obtain an algorithm for the n to 1 linear transformation for the unequal mean and covariance two-class Gaussian data. In Section 5, we will consider some constrained linear transformations for sequentially received data, using the Bhattacharyya distance. In Section 6, we consider and discuss the situation in which the data can be modeled by Gauss-Markov processes. In Section 7, we consider briefly the special case when the covariance matrices are Toeplitz. In Section 8, conclusions and remarks will be offered.
2. Preliminaries

Let \((\Theta, W, F)\) be a separable, complete measure space, \(W\) is a \(\sigma\)-algebra of subsets of \(\Theta\) and \(F\) is the measure. Let \(F_1\) and \(F_2\) be two different measures in \((\Theta, W)\) such that \(F_1(\Theta) = F_2(\Theta)\). Suppose also that \(p(\cdot, \cdot)\) is a function called penalty or distortion function defined on sets in \(W\), then the Prokhorov and Vasershtein distances denoted by \(D_P\) and \(D_V\) are defined as follows:

\[
D_P(F_1, F_2) = \inf \{ \epsilon \in \mathbb{R} : F_1(A) \leq F_2(U \cup B : p(A, B) \leq \epsilon) + \epsilon, \\
F_2(A) \leq F_1(U \cup B : p(A, B) \leq \epsilon) + \epsilon ; \\
\forall A \in W \}\]  

(2.1)

\[
D_V(f_1, f_2) = \inf_{A, B \in W} E_{F(\cdot, \cdot)} \{ p(A, B) \}
\]

(2.2)

where \(F(\cdot, \cdot)\) is the joint measure with \(F_1(\cdot)\), \(F_2(\cdot)\) as the marginals. We see that the Prokhorov distance is bounded from above by \(F_1(\Theta)F_2(\Theta)\), and that the boundedness of \(D_V(\cdot, \cdot)\) depends on the function \(p(A, B)\). Dobrushin [52] showed that if \(p(\cdot, \cdot)\) is a metric, then \(D_V(\cdot, \cdot)\) is also a metric. The Prokhorov distance, however, is a metric for any choice of \(p(\cdot, \cdot)\). Kazakos [53] showed that

\[
D_V(\cdot, \cdot) \geq (D_P(\cdot, \cdot))^2
\]

(2.3)
The Lévy distance is a special case of the Prokhorov distance when \( \Theta \) is Euclidean and the measures \( F_1(\cdot) \), \( F_2(\cdot) \) are distribution functions. When \( \Theta \) is the real line, the Levy distance \( D_L \) is given by:

\[
D_L(F_1, F_2) = \inf \{ \varepsilon : F_2(\min(y : p(x,y) \leq \varepsilon)) - \varepsilon \leq F_1(x) \\
\leq F_2(\max(y : p(x,y) \leq \varepsilon)) + \varepsilon, \forall x \in \Theta \}
\]

(2.4)

The Lévy distance is also a metric. We list below five distances that are defined on Euclidean spaces:

Kolmogorov:

\[
D_K(F_1, F_2) = \sup_x |F_1(x) - F_2(x)|
\]

(2.5)

Variational:

\[
D_{VR}(f_1, f_2) = \int |f_1(x) - f_2(x)| \, dx
\]

(2.6)

I-Divergence:

\[
D_D(f_1, f_2) = \frac{1}{2} \int (f_1(x) - f_2(x)) \log \frac{f_1(x)}{f_2(x)} \, dx
\]

(2.7)

Chernoff:

\[
D_C(f_1, f_2, t) = -\ln \int f_1^t(x) f_2^{1-t}(x) \, dx \quad t \in [0, 1]
\]

(2.8)

Bhattacharyya:

\[
D_B(f_1, f_2) = D_C(f_1, f_2, 1/2)
\]

(2.9)
The functions $f_1$, $f_2$ are density functions corresponding to the distribution functions $F_1$, $F_2$, respectively. The region of integration in (2.6-2.9) is the sample space of $x$. Dobrushin [52] showed that for particular choices of $p(\cdot,\cdot)$, both the Prokhorov and Vasershtein distances degenerate to the Kolmogorov distance.

**Lemma 2.1**

If $F_1$ and $F_2$ are distribution functions, and $f_1$, $f_2$ the corresponding density functions, then

$$D_{\pi}(F_1, F_2) \leq D_K(F_1, F_2) \leq D_{VR}(f_1, f_2) \quad (2.10)$$

**Proof:** See [53].

**Lemma 2.2**

If the functions $f_1$ and $f_2$ are the class conditional densities, and if the two priors $q_1$, $q_2$ are equal, then

$$D_{VR}(f_1, f_2) = 2(1 - 2 P_e) \quad (2.11)$$

**Proof:**

$$P_e = 1 - \int \max (q_1 f_1(x), q_2 f_2(x)) dx$$

$$= 1/2 \left[ 1 - \int |q_1 f_1(x) - q_2 f_2(x)| dx \right]$$

$$= 1/2 \left[ 1 - \frac{1}{2} \int |f_1(x) - f_2(x)| dx \right]$$

$$= 1/2 \left[ 1 - \frac{1}{2} D_{VR}(f_1, f_2) \right]$$
where the integration region is the sample space of \( x \).

We see that for the equal priors case, knowing the Variational distance is equivalent to knowing the probability of error. Hence, we can expect similar computational difficulties in using \( D_{VR}(\cdot,\cdot) \), for feature selection purposes, as those encountered in using \( P_e \).

As mentioned earlier, in this part of the thesis, we will concentrate on the Bhattacharyya distance which is a special case of the Chernoff distance. Kailath [54] using a theorem attributed to Blackwell [55], showed that if for two sets of parameters \( a_1 \) and \( a_2 \) such that \( D_{B,a}(\cdot,\cdot) \) is greater than \( D_{B,a_2}(\cdot,\cdot) \), then there exists at least one set of priors \( q_1 \) and \( q_2 \) for which \( P_e(a_1,q_1,q_2) \) is less than \( P_e(a_2,q_1,q_2) \). A similar property exists for the divergence [56]. Kailath also showed that the Bhattacharyya distance bounds the probability of error in the following manner:

\[
q_1 - \frac{1}{2} \left( 1 - 4q_1q_2 \exp(-2D_B) \right)^{\frac{1}{2}} \leq P_e \leq q_1 + \left( q_1q_2 \right)^{\frac{1}{2}}
\]

\[
\exp(-D_B) - \frac{1}{2} \leq P_e \leq \frac{1}{2} \exp(-D_B)
\]

which reduces to

\[
\frac{1}{4} \exp(-2D_B) \leq P_e \leq \frac{1}{2} \exp(-D_B)
\]
when \( q_1 \) equals \( q_2 \). Although several distances such as the Ito's measure \([57]\) and the Bayesian distance \([48]\) have tighter bounds on \( P_e \) than the Bhattacharyya distance, their calculabilities are generally much more difficult. An advantage of the Bhattacharyya distance is that closed form expressions for exponential densities can be found, thus making it more amenable to analysis.

In the following section, linear transformations will be considered. Special cases of Gaussian statistics will be used to obtain these transformations using the distances described above.
3. **Linear Transformations**

Since we are interested in transforming finite dimensional probability spaces, let us consider two measure spaces \( (\Theta, \mathcal{W}, \nu_1), (\Theta, \mathcal{W}, \nu_2) \), and a linear transformation \( T \). Let the transformed space be denoted by \( T\Theta \), and the transformed \( \sigma \)-algebra \( \mathcal{W} \) by \( T\mathcal{W} \). We note that \( T\mathcal{W} \) will also be a \( \sigma \)-algebra if \( \Theta \) is countable. Consider now a closed convex set \( T^0 \) of transformations on \( (\Theta, \mathcal{W}, \nu) \) such that the transformed \( \sigma \)-algebra is also a \( \sigma \)-algebra, then each \( T \in T^0 \) induces a unique measure space \( (T\Theta, T\mathcal{W}, f_T(v)) \). If a distance of the measures \( f_T(v_1) \) and \( f_T(v_2) \), denoted by \( D(f_T(v_1), f_T(v_2)) \), is concave with respect to \( T \in T^0 \), then there is a unique transformation \( T^* \in T^0 \) such that \( D(f_{T^*}(v_1), f_{T^*}(v_2)) \) is maximized.

In this section, the set \( \Theta \) will be the \( n \)-dimensional Euclidean space, \( \mathbb{E}^n \), and the \( \sigma \)-algebra \( \mathcal{W} \) will be the \( n \)-dimensional intervals \( (-\infty, x) \) where \( x \in \mathbb{E}^n \). The set \( T^0 \) will be the family of row vectors, \( t \), satisfying the restriction \( t^R t^T = C \), where \( R \) is a positive definite matrix. The measures \( \nu_1, \nu_2 \) will be probability measures. Specifically, they will be Gaussian probability measures defined on \( \mathcal{W} \). Hence,
\[ v_i(-\infty, x) = \int_{-\infty}^{x} \frac{1}{(2\pi)^{n/2}} \left| R_i^{-1} \right|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2}(y-m_i)^T R_i^{-1} (y-m_i) \right\} \, dy \]

where \( m_1, m_2 \) are the vector means and \( R_1, R_2 \) the covariance matrices. Thus, for \( T = t \), the induced measure, \( f_T(v_i) \), is obtained by replacing \( m_i \) by \( tm_i t^T \) and by \( tR_i t^T \) in (2.14). The constraint on \( t \) will be

\[ tr_i t^T = C \]

where \( C \) is a positive constant.

We will now use several of the distances as criterion functionals in computing the optimal transformation \( t^* \) for two specific cases. Firstly, we will consider the case where \( m_1 = m_2 = m \). The equal covariance case where \( R_1 = R_2 = R \) will then be considered. Before we proceed, a brief digression on the generalized characteristic equation is presented to establish some facts and notations.

Consider the following generalized characteristic equation:

\[ tR_2 = \lambda tR_1 \]

where \( t \) is a row vector of dimension \( n \), the matrix \( R_1 \) is positive definite, and the matrix \( R_2 \) is symmetric. Thus, for two distinct eigenvalues \( \lambda_1, \lambda_2 \) and their corresponding eigenvectors \( t_1, t_2 \) we have:
\[ (\lambda_1 - \lambda_2) \, t_1 R_1 t_2^T = 0 \]  

(3.4)

indicating that \( t_1 \) and \( t_2 \) are orthogonal relative to \( R_1 \).

When \( \lambda_1 = \lambda_2 \) then the Gram-Schmidt orthogonalization method can be used to orthogonalize \( t_1 \) and \( t_2 \). Thus, the eigenvectors of \((3.3)\) can be taken to be orthonormal. relative to \( R_1 \). Let the matrix \( L^{-1} \) be the modal matrix:

\[
L^{-1} = \begin{bmatrix}
    t_1 \\
    \vdots \\
    t_n
\end{bmatrix}
\]  

(3.5)

where each \( t_i, i=1,n \) is a row vector and \( t_i R_1 t_j^T \) is the Dirac delta function. Equation \((3.3)\) can now be written in the matrix form:

\[
L^{-1} R_2 = D \, L^{-1} R_1
\]  

(3.6)

where \( D \) is the diagonal matrix with the eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_n \) as the diagonal elements. We have also that

\[
R_1 = L \, L^T
\]  

(3.7)

\[
R_2 = L \, D \, L^T
\]  

(3.8)

**Definition 3.1**

We will denote the eigenvector \( t_1 \) that corresponds to the largest value of \( \lambda_1 + 1/\lambda_1 \) by \( t(1) \), and the next largest by \( T(2) \), and so on. For ordered eigenvalues such
that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, then $t(1) = t_1$ or $t_n$.

**Definition 3.2**

Similarly, we will denote $\lambda(1)$ to be the eigenvalue $\lambda_i$ such that $\lambda_i + 1/\lambda_i$ gives the largest value. The eigenvalue $\lambda_j$ with the smallest value of $\lambda_j + 1/\lambda_j$ will be denoted by $\lambda(n)$.

**3.1 Equal Covariance Case: $R_1 = R_2 = R$**

**A. Bhattacharyya Distance**

The Bhattacharyya distance in this case becomes:

$$D_B(f_t(v_1), f_t(v_2)) = \frac{1}{8} (tm)^2/(tRt^T)$$

(3.9)

which using (3.2) implies that

$$t^* = \pm \frac{1}{2} \sqrt{\frac{R}{m}} mR^{-1}$$

(3.10)

maximizes the value of $D_B(\cdot, \cdot)$. At $t^*$ the value of the Bhattacharyya distance is:

$$D_B, t^* = 1/8 \ m^T R m$$

(3.11)

We should mention here that (3.11) is the maximum value achievable in any valid dimension.

**B. Variational Distance**

The variational distance can be expressed by:
\[ D_{VR}(f_t(v_1), f_t(v_2)) = 2 \left| \varphi \left( \frac{tm}{2(tR + T)^{\frac{3}{2}}} \right) - \varphi \left( -\frac{tm}{2(tR + T)^{\frac{3}{2}}} \right) \right| \quad (3.12) \]

where

\[ \varphi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{u^2}{2} \right] \, du \quad (3.13) \]

The maximum value of (3.12) is achieved when \( \left| \frac{tm}{2(tR + T)^{\frac{3}{2}}} \right| \) is maximized subject to (3.2). This of course gives the \( t^* \) as in (3.10). From Lemma 2.2 we see that this \( t^* \) achieves the minimum \( P_e \). The value of \( D_{VR} \) at \( t^* \) is given by:

\[ D_{VR,t^*} = 2 \left[ \varphi \left( (mR + T)^{\frac{3}{2}}/2 \right) - \varphi \left( -(mR + T)^{\frac{3}{2}}/2 \right) \right] \quad (3.14) \]

C. Kolmogorov Distance

In this case:

\[ D_K(f_t(v_1), f_t(v_2)) = 1/2 \, D_{VR}(f_t(v_1), f_t(v_2)) \quad (3.15) \]

and \( t^* \) is as in (3.10).

D. \( D \)-Divergence

It is easy to find that

\[ D_D(f_t(v_1), f_t(v_2)) = 1/4 \, D_B(f_t(v_1), f_t(v_2)) \quad (3.16) \]

Hence, once again, the \( t^* \) is given by (3.10).

E. Lévy Distance

The Lévy distance can be expressed by (see [53]):
\[ D_L(f_t(v_1), f_t(v_2)) = \inf \{ \varepsilon : g\left( \frac{-p^{-1}(\varepsilon)}{2(tRT)^{\frac{1}{2}}} \right) \} - \]

\[ g\left( \frac{-tm}{2(tRT)^{\frac{1}{2}}} \right) \leq \varepsilon \]  \hspace{1cm} (3.17)

where

\[ p^{-1}(\varepsilon) = \max (y : p(x,x+y) = \varepsilon \forall x) \]  \hspace{1cm} (3.18)

We see that the condition \( tRT^T = C \) implies that \( t^* \) is again given by (3.10).

F. Vasershtein Distance

Suppose that the \( n \) dimensional Gaussian data \( x \) are generated by first order wide sense stationary Markov processes, then we can find the power spectra for \( tx \) under both classes. Specifically, the power spectra \( S_1(\lambda) \) and \( S_2(\lambda) \) are given by:

\[ S_1(\lambda) = \sum_{k=-\infty}^{\infty} tR_1(k)t^T \exp \{ jk\lambda \} \]  \hspace{1cm} (3.19)

\[ S_2(\lambda) = \sum_{k=-\infty}^{\infty} tR_2(k)t^T \exp \{ jk\lambda \} \]  \hspace{1cm} (3.20)

where \( R_1(k) = E \left[ (x(n)-m_1(n))(x(n+k)-m_1(n+k))^T \right] \)  \hspace{1cm} (3.21)

and \( x(n) \) denotes the data \( x \) at time \( n \).

For \( p(x,y) = (x-y)^2 \), we can show that

\[ D_V(f_t(v_1), f_t(v_2)) = (2\pi)^{-1} \int_{-\pi}^{\pi} \left| S_1(\lambda) - S_2(\lambda) \right| 2d\lambda + (tm)^2 \]  \hspace{1cm} (3.22)
since $S_1(\lambda) = S_2(\lambda)$, the Vaserstein distance becomes:

$$D_V(f_t(v_1), f_t(v_2)) = (tm)^2 \quad (3.23)$$

Thus (3.23) is maximized when $t^*$ is, again, given by (3.10).

### 3.2 Equal Mean Case: $m_1 = m_2 = m$

#### A. Bhattacharyya Distance

The Bhattacharyya distance for this case is given by:

$$D_B(f_t(v_1), f_t(v_2)) = \ln\left(\frac{\frac{1}{tR_1 t^T} \frac{1}{tR_2 t^T}}{\frac{1}{tR_1 t^T} \frac{1}{tR_2 t^T}}\right)^{\frac{1}{2}} + \frac{1}{2} \left(\frac{tR_2 t^T}{tR_1 t^T}\right)^{\frac{1}{2}} \quad (3.24)$$

Using (3.7) we can easily obtain the optimal transformation:

$$t^* = C^{\frac{1}{2}} t_{(1)} \quad (3.25)$$

where $t_{(1)}$, is defined in Definition 3.1.

#### B. Variational Distance

In this case,

$$D_V(f_t(v_1), f_t(v_2)) = 4 \left| \begin{array}{c}
\phi \left( \frac{tR_2 t^T}{tR_1 t^T} \right) \frac{2 \ln \frac{tR_2 t^T}{tR_1 t^T}}{(\frac{2}{tR_2 t^T}) - 1} \\
\phi \left( \frac{tR_2 t^T}{tR_1 t^T} \right) \frac{2 \ln \frac{tR_2 t^T}{tR_1 t^T}}{(\frac{2}{tR_2 t^T}) - 1} \\
\end{array} \right| \quad (3.26)$$
and (3.25) gives the optimal transformation.

C. Kolmogorov Distance

It can be shown that $D_K(\cdot, \cdot) = \frac{1}{4} D_{VR}(\cdot, \cdot)$ (3.27)
and so (3.25) is the optimal transformation.

D. I-Divergence

We easily obtain that

$$D_D(f_t(v_1), f_t(v_2)) = -\frac{1}{2} + \frac{1}{4} \left( \frac{tR_2 t^T}{tR_1 t^T} + \frac{tR_1 t^T}{tR_2 t^T} \right)$$ (3.28)

and hence, $t^* = C^{\frac{1}{2}} t(1)$ maximizes (3.28).

E. Lévy Distance

The Lévy distance for this case is expressed by:

$$D_L(f_t(v_1), f_t(v_2)) = \inf \{ \varepsilon : \wp(y) - \wp(y + \frac{ytR_1 t^T}{tR_2 t^T} + \frac{p^{-1}(\varepsilon)}{tR_2 t^T}) \leq \varepsilon ;$$
$$\wp \left( \frac{ytR_1 t^T}{tR_2 t^T} - \wp(y + \frac{p^{-1}(\varepsilon)}{tR_2 t^T}) \right) \leq \varepsilon ;$$
$$\forall y \}$$ (3.29)

where $p^{-1}(\varepsilon)$ is given by (3.18).

The computation involved in (3.29) can only be done numerically. We will not attempt to find $t^*$ for this distance.
F. Vasershtein Distance

If we consider wide sense stationary processes as for the equal covariance case, the Vasershtein distance with \( p(x,y) = (x-y)^2 \) becomes (see [58]):

\[
D_V(f_t(v_1), f_t(v_2)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| S_1^{\frac{1}{2}}(\lambda) - S_2^{\frac{1}{2}}(\lambda) \right|^2 d\lambda \tag{3.30}
\]

where \( S_1(\lambda) \) and \( S_2(\lambda) \) are given by (3.19) and (3.20), respectively. For first order Markov processes (3.19) and (3.20) become:

\[
S_1(\lambda) = tr_1(0)t^T + 2tr_1(1)t^T\cos \lambda \tag{3.31}
\]

\[
S_2(\lambda) = tr_2(0)t^T + 2tr_2(1)t^T\cos \lambda \tag{3.32}
\]

which when substituted in (3.30) gives:

\[
D_V(f_t(v_1), f_t(v_2)) = t \left[ R_1(0) + R_2(0) \right] t^T -
\]

\[
\frac{1}{2\pi} \int_{-\pi}^{\pi} (t(R_1(0) + 2R_1(1)\cos \lambda)t^T \cdot t(R_2(0) + 2R_2(1)\cos \lambda)) \cdot t^T)^{\frac{1}{2}} d\lambda \tag{3.33}
\]

Again, (3.33) is difficult to evaluate analytically. By applying the Schwartz inequality to (3.33), we obtain:

\[
D_V(f_t(v_1), f_t(v_2)) \geq ((tR_1(0)t^T)^{\frac{1}{2}} - (tR_2(0)t^T)^{\frac{1}{2}})^2 \tag{3.34}
\]

where equality holds if and only if there exists a
constant $A$ such that

$$t( (R_1(0) - AR_2(0)) + 2(R_1(1) - AR_2(1))\cos \lambda) t^T = 0 \quad (3.35)$$

for almost all $\lambda$ in $[-\pi, \pi]$. Maximizing (3.34) instead, subject to (3.2), gives the same $t^*$ as in (3.25). We emphasize here that the Vasershstein distance, as well as the lower bound, depend only on the second order statistics of the data. As a result, this distance is less sensitive to the variations in the underlying probability density functions.
4. Bhattacharyya Distance; Unequal Mean and Covariance

For Gaussian densities with unequal mean and covariance, the Bhattacharyya distance is expressed by:

\[ D_B = \frac{1}{4} \left( m^T (R_1 + R_2)^{-1} m + \ln \frac{1}{2} \left( R_1 + R_2 \right) \right)^2 \]

\[ \frac{1}{|R_1|} \frac{1}{|R_2|} \]  \hspace{1cm} (4.1)

If we consider a linear transformation \( t \), of dimension \( mxn \), where \( n \) is the dimensionality of the data vector and \( m = n \), then the Bhattacharyya distance in \( \mathbb{E}^m \) is given by:

\[ D_{B,t} = \frac{1}{4} \left( m^T t^T (t(R_1 + R_2) t^T)^{-1} t m + \right. \]

\[ \left. \ln \frac{1}{|t^T R_1 t|^2} \frac{1}{|t^T R_2 t|^2} \right) \]  \hspace{1cm} (4.2)

We observe that if \( Q \) is any positive definite \( mxm \) matrix, then

\[ D_{B,t} Q t = D_{B,t} \]  \hspace{1cm} (4.3)

As a result of (4.3), the optimal rank \( m \) transformation \( t^* \) can be chosen to be such that

\[ t^* t^{*T} = I_m \]  \hspace{1cm} (4.4)

**Lemma 4.1**

A solution to the problem
\[
\max_{\{t\}} D_B, t \text{ such that } t t^T = I_m, \ t \in \mathbb{R}^{m \times n}
\]
exists.

**Proof:**

The constraint \( t t^T = I_m \) implies that we need to research for \( t^* \) on the surface of an \( n \) dimensional hypersphere which is a compact set. Since the function \( D_B, t \) is continuous on the hypersphere, we conclude that extrema exist.

From (4.2) we observe that for \( t \in \mathbb{R}^{m \times n} \):

\[
D_B, t^* \leq 1/4 \left( \max_{\{t\}} m t^T (t(R_1 + R_2)t^T)^{-1} t m + \right.
\]

\[
\max_{\{t\}} \ln \left. \frac{\frac{1}{2} t(R_1 + R_2)t^T}{|tR_1^T| |tR_2^T|} \right|^{\frac{2}{}} \right)
\]

(4.5)

Using the notations defined in Definition 3.2, and the equations (3.3-3.8), the above expression (4.5) becomes:

\[
D_B, t^* \leq 1/4 \left( m^T (R_1 + R_2)^{-1} m + \sum_{i=1}^{m} \ln \frac{1}{2} (\lambda(i) + \frac{1}{\lambda(i)}) + 2 \right)
\]

(4.6)

where equality is achieved if and only if \( m^T (R_1 + R_2)^{-1} \) is in the row space of the eigenvectors \( t(1), t(2), \ldots, t(m) \).

We note that since \( R_1 \) and \( R_2 \) are assumed to be positive definite, the eigenvalues are all positive. Also, the
minimum value of $\lambda + 1/\lambda$, where $\lambda$ denotes an eigenvalue of the generalized characteristic equation (3.3), occurs when $\lambda = 1$.

A numerical method to compute $t^*$ by setting the gradients of (4.2) to zero was presented in [47]. In this section an algorithm for the transformation of n dimensional data to a scalar will be presented. It is hoped that the specific nature of the problem would result in a more practical algorithm. One advantage of our algorithm is its adaptability to parallel processing.

We now consider the problem posed in Lemma 4.1 with $t$ now $\in \mathbb{R}^{1 \times n}$ and $t^T t = 1$.

For $t \in \mathbb{R}^{1 \times n}$, (4.2) can be rewritten as:

$$D_{B,t} = \frac{1}{4} \left( \frac{(tm)^2}{t(R_1+R_2)^T t^T} + \ln \frac{1}{4} \frac{(t(R_1+R_2)t^T)^2}{(tR_1 t^T)(tR_2 t^T)} \right)$$

(4.7)

Applying the transformation

$$y = \frac{\frac{tL}{\| v \|}}$$

(4.8)

to (4.7), we obtain:

$$D_{B,y} = \frac{1}{4} \left( -\frac{(yA)^2}{1+yDy^T} + \ln \ln \frac{1}{4}(yDy^T+1/yDy^T+2) \right)$$

(4.9)

where

$$A = L^{-1} M$$

(4.10)

and $L$ is given by (3.5).
Since $y$ is restricted to have unit norm, we also have that:

$$\lambda_1 \leq yDy^T \leq \lambda_n \quad (4.11)$$

where the $\lambda$'s have been ordered.

Thus, the problem of finding the $t^*$ is equivalent to finding a $y^*$ such that $D_B, y^*$ is a maximum, and that $y^*$ has unit norm. Recognizing that if we momentarily fix the value of $yDy^T$ to be somewhere in the interval $[\lambda_1, \lambda_n]$, then we can always find unit vectors $y$. In other words, the unit norm constraint and the constraint in (4.11) are not disjoint. The problem, therefore, is equivalent to searching over the allowable values of $yDy^T$ to locate a $y^*$ that maximizes (4.9). We note that this is similar to solving the following problem,

$$\max_{y} (yA)^2 \text{ such that } yy^T = 1, \quad yDy^T \in [\lambda_1, \lambda_n] \quad (4.12)$$

We will denote the momentarily fixed value of $yDy^T$ by $Z$. We further note that any $y$ that extremizes $(yA)$ will also maximize $(yA)^2$. Thus, our final problem is to solve:

Extremize $yA$ such that $yy^T = 1, \quad yDy^T = Z \quad (4.13)$

To solve (4.13), let us use Lagrange multiplier method with multipliers $u_1$ and $u_2$. Let the Lagrange function be denoted by $LG(\cdot)$. then
\[ \text{LG}(y) = yA - u_1 yy^T - u_2 yDy^T \] \hspace{1cm} (4.14)

Taking the gradient of (4.14) and setting it to zero, we obtain:

\[ y = \frac{A^T}{2} (u_1 I + u_2 D)^{-1} \] \hspace{1cm} (4.15)

Substituting (4.15) into (4.13) gives:

\[ \sum_{i=1}^{n} \frac{a_i^2}{(u_1 + u_2 \lambda_1)^2} = 4 \] \hspace{1cm} (4.16)

and

\[ \sum_{i=1}^{n} \frac{a_i^2 \lambda_1}{(u_1 + u_2 \lambda_1)^2} = 4Z \] \hspace{1cm} (4.17)

where the \( a_i \)'s are the components of the vector \( A \), and the \( \lambda_1 \)'s are the elements of the diagonal matrix \( D \).

Also, substituting (4.15) into (4.9), we can re-express the Bhattacharyya distance as:

\[ D_B(u_1(Z),u_2(Z),Z) = \frac{1}{4} \left[ \frac{(A^T(u_1(Z)I+u_2(Z)D)^{-1}A)^2}{4(1+Z)} + \ln \left( \frac{(Z + 1/Z + 2)/4}{(Z + 1/Z + 2)/4} \right) \right] \] \hspace{1cm} (4.18)

Combining (4.16) and (4.17), we obtain:

\[ \sum_{i=1}^{n} \frac{(z-\lambda_1)a_i^2}{(1+u_i^2 \lambda_1)^2} = 0 \] \hspace{1cm} (4.19)

where
u = u_2 / u_1 \quad (4.20)

and \( z \in [\lambda_1, \lambda_n] \).

We observe from (4.19) that if all the \( a_i \)'s, except for \( a_j \), are equal to zero, then \( z \) must be equal to \( \lambda_j \). From the constraints in (4.13), we see that this implies that the optimal transformation \( t^* \) is equal to some nonzero constant multiple of the eigenvector \( t_j \). We note that for \( t^* = t_1 \) or \( t_n \), the condition on the \( a_i \)'s in the above observation must hold. We emphasize here that the procedure developed in this section assumes that \( A \neq 0 \). If \( A = 0 \), then we must use (4.9) to obtain the optimal transformation. It is not difficult to see from (4.9) that if \( A = 0 \), then \( t^* = t_{(1)} \) as in the equal mean case.

We list below the conditions under which the \( t^* \) can be obtained without further calculation:

Case 1 \( m = 0 \) or \( A = 0 \) & \( R_1 \neq C R_2 \Rightarrow t^* = kt_{(1)} \)

Case 2 \( a_j = 0, a_{i \neq j} = 0 \) & \( R_1 \neq C R_2 \Rightarrow t^* = kt_j \)

Case 3 \( m \neq 0 \) & \( R_1 \neq C R_2 \Rightarrow t^* = km^{T_R} \)

where \( C \) and \( k \) are some nonzero constants.

We now consider the general case in which (4.19) must be computed. The procedure we will follow is to search over the dense interval \( (\lambda_1, \lambda_n) \) for the value of
Z that will maximize the Bhattacharyya distance (4.18). To obtain the values of \( u \), we can solve the polynomial equation contained in (4.19). We describe below our procedure for obtaining the coefficients of the polynomial. Before doing so, a brief digression on symmetric functions is presented below.

4.1 Symmetric Functions

We will deal with elementary symmetric functions which are defined below.

**Definition 4.1**

Given \( N \) numbers \( x_1, x_2, \ldots, x_N \) where \( N \geq 1 \), the functions

\[
\begin{align*}
  f_0(x_1, x_2, \ldots, x_N) &= 1, \\
  f_1(x_1, x_2, \ldots, x_N) &= \sum_{i=1}^{N} x_i, \\
  f_2(x_1, x_2, \ldots, x_N) &= \sum_{1 \leq i < j \leq N} x_i x_j, \\
  f_m(x_1, x_2, \ldots, x_N) &= \sum_{1 \leq i_1 < i_2 < \ldots < i_m \leq N} x_{i_1} x_{i_2} \cdots x_{i_m}, \\
  f_N(x_1, x_2, \ldots, x_N) &= \prod_{i=1}^{N} x_i,
\end{align*}
\]

are called the elementary symmetric functions of the \( N \) objects.
We will establish in this section an iterative expression for the \( N \) elementary symmetric functions. For brevity the \( j \)th elementary symmetric function will be denoted by \( f_j \).

**Lemma 4.2**

Let \( ! \) denote the factorial, then,

\[
f_1 = (K+1)! f_{K+1} - \left( f_1 g_2 + 2! f_1 g_3 + \cdots + K! g_{K+1} \right) \tag{4.21}
\]

where

\[
g_K = \sum_{i=2}^{K} (-1)^{i+1} h_i f_{i-1}^{K-1} \tag{4.22}
\]

and

\[
h_i = \sum_{j=1}^{N} x_j^i \tag{4.23}
\]

**Proof:**

We will prove (4.21) by induction.

For \( K = 1 \), the right hand side of (4.21) is \( 2f_2 + h_2 \) which is equal to the left hand side.

Assume (4.21) is true for \( K = L \), then

\[
f_{L+1} = f_1 f_L
\]

\[
= f_1 \left[ L! f_L - (f_1 g_2 + \ldots + (L-1)! g_{L-1}) \right]
\]

Now \( f_1 f_L \) is a sum of terms that can be grouped into two types as long as \( N > L \). Type 1 consists of terms that
are a product of $L + 1$ numbers $x_{i1}x_{i2}\ldots x_{iL+1}$ where the indices are all distinct. Type 2 consists of terms in which two of the indices are the same. For example, if $L = 2$, a type 1 term for $f_1 f_2$ is $x_1 x_2 x_3$ and a type 2 term is $x_1^2 x_2$. We know that there are $\binom{N}{L+1} \cdot (L+1)$ type 1 terms out of $\binom{N}{L} \cdot N$ total number of terms. A type 2 term can be considered as a type 1 term of length $L$ multiplied by a squared number. The length $L$ type 1 terms associated with $x_{im}^2$ necessarily do not contain $x_{im}$. If we do include the number $x_{im}$ in the length $L$ type 1 terms, then the type 2 terms for $f_1 f_L$ can be expressed by $h_2 f_{L-1}^2$ (terms involving cubed numbers). If we follow the same logic of adding and subtracting the extra term, the parts of $f_1 f_L$ that involve the type 2 terms can be expressed by:

$$h_2 f_{L-1}^2 - h_3 f_{L-2}^2 + h_4 f_{L-3}^2 - \ldots (-)^{L+1} h_L f_1 + (-1)^{L+1} h_{L+1}.$$ 

Thus, $f_1 f_L = (L+1) f_{L+1} + h_2 f_{L-1}^2 - h_3 f_{L-2}^2 + \ldots (-1)^{L+1} h_{L+1} f_{L-1}^2$.

Hence, $L! f_1 f_L = (L+1)! f_{L+1} - L! \sum_{j=2}^{L+1} (-1)^{j+1} h_j f_{L-j+1}$.

Therefore $f_1^{L+1} = (L+1)! f_{L+1} - (f_1^{L-1} g_2 + \ldots +(L-1)! \cdot f_1 g_L + L! g_{L+1})$ which is of the form in (4.21).
Lemma 4.3

The Kth elementary function can be expressed by:

\[ f_K = \frac{1}{K} (f_1f_{K-1} + \sum_{j=2}^{K} (-1)^{j+1} h_j f_{K-j} ) \]  \hspace{1cm} (4.24)

where \( h_j \) is given by (4.23)

Proof:

For \( K=1 \) and 2 (4.24) is obviously true. Let us assume that (4.24) is true for \( K=L \). For \( K=L+1 \) then,

\[
f_{L+1} = \frac{1}{L+1} \left( f_1f_L + \sum_{j=2}^{L+1} (-1)^{j+1} h_j f_{L+1-j} \right) \\
= \frac{1}{L+1} \left[ f_1 \frac{1}{L} \left( f_1f_{L-1} + \sum_{j=2}^{L} (-1)^{j+1} h_j f_{L-j} \right) + \right. \\
\left. \sum_{j=2}^{L+1} (-1)^{j+1} h_j f_{L+1-j} \right] \\
= \frac{1}{L+1} \left[ \left( \frac{1}{L} \right)^2 \left( f_1^2 f_{L-2} + \sum_{j=2}^{L} (-1)^{j+1} h_j f_{L-1-j} \right) + \right. \\
\left. \sum_{j=2}^{L+1} (-1)^{j+1} h_j f_{L+1-j} \right] \\
\]

Using (4.22),

\[
f_{L+1} = \frac{1}{L+1} \left\{ \frac{1}{L} \left[ \frac{1}{L-1} (f_1^3 f_{L-2} + f_1^2 g_{L-1}) + f_1 g_L \right] + g_{L+1} \right\} \\
\]

which if we continue becomes:

\[
f_{L+1} = \frac{1}{(L+1)!} \left( f_1^{L+1} + f_1 L-1 g_2 + 2! f_1^{L-2} g_3 + \ldots + L! g_{L+1} \right) \\
\]

From Lemma 4.2 the above equation is true and hence (4.24)
holds.

From (4.23) we see that each $h_i$ for $i=2,3,...,N$ requires $N$ multiplications. Thus, the total number of multiplications required to compute the $K$th elementary symmetric function $K < N$, from (4.24) is $N + K - 1$ which includes the division by $K$. The number of adds is $N + K - 2$.

Varti [59] had reported that the $K$th elementary symmetric function can be computed with $(K-1)(N-K+1)$ multiplications, and $K(N-K)$ additions. A comparison of the number of multiplications can be made by finding conditions on $K$ and $N$ such that:

$$(K-1)(N-K+1) \geq N+K-1 \quad (4.25)$$

From (4.25) it is not difficult to see that for

$$K \leq \min \text{ integer part } \left( \frac{1 + N^2(N^2 - 6N + 1)}{2} \right) \quad (4.26)$$

and that

$$N \geq 6 \quad (4.27)$$

the iterative scheme requires less number of multiplications. A similar calculation on the number of additions shows that for $N \geq 7$ and $K \leq \min$ integer part

$$\left( \frac{N-1 + (N^2 - 6N - 7)}{2} \right)^{\frac{1}{3}}$$

the iterative scheme requires less number of additions. The number of multiplications and
additions in [59] is based on the decomposition scheme presented below.

**Lemma 4.4**

For any \( L \) such that \( 1 \leq L < N \), and for any numbers \( x_1, x_2, \ldots, x_N \) which are elements of an arbitrary ring:

\[
f_K(x_1, x_2, \ldots, x_N) = \begin{cases} 
\sum_{j=0}^J f_{K-j}(x_1, \ldots, x_L) \cdot f_j(x_{L+1}, \ldots, x_N) \\
\sum_{j=K-M}^L f_{K-j}(x_1, \ldots, x_L) \cdot f_j(x_{L+1}, \ldots, x_N) 
\end{cases}
\]

where \( J = \min(K, N-L) \).

**Proof:**

See Goyal [60].

The decomposition in lemma 4.4 has also been used as the basis for a parallel computational scheme to compute the elementary symmetric functions [60]. We mention here that the computational complexity depends on the underlying space in which the numbers exist. Strassen [61] has shown that for closed complex fields the lower bound on the number of multiplications is \( O(N \log_2 N) \). Over fields in which the Fast Fourier Transform is
meaningful, Goyal mentioned that Papadimitriou obtained a parallel algorithm requiring $O((\log_2 N)^2)$ steps using $O(N(\log_2 N)^2)$ operations to compute the elementary symmetric functions. Over an arbitrary ring, Goyal [60] has presented a parallel scheme requiring $O((\log_2 N)^2)$ steps, using $O(N^2)$ operations.

This concludes our brief digression into the elementary symmetric functions. We return now to see how they are used in computing (4.19).

### 4.2 An Algorithm

Equation (4.19) can be expressed as follows:

$$\sum_{i=1}^{n} (z - \lambda_i) a_i^2 \prod_{j \neq i} (1 + u \lambda_j)^2 = P(u) = 0$$  \hspace{1cm} (4.29)

Let

$$Q(n) = \prod_{j \neq i} (1 + u \lambda_i)$$  \hspace{1cm} (4.30)

and

$$Q_j(n) = \frac{\prod_{i=1}^{n} (1 + u \lambda_i)}{(1 + u \lambda_j)} = \frac{Q(n)}{(1 + u \lambda_j)}$$  \hspace{1cm} (4.31)

then

$$Q_j(n)(1 + u \lambda_j) = Q(n)$$  \hspace{1cm} (4.32)

which when expanded becomes:

$$\sum_{i=0}^{n-1} C_i^j u^i + \sum_{i=1}^{n} \lambda_j C_i^{i-1} u^i = \sum_{i=0}^{n} d_i u^i$$  \hspace{1cm} (4.33)
where $C^j_i$ and $d_i$ denote the coefficients of the polynomials $Q^j(n)$ and $Q(n)$, respectively. Grouping the coefficients of like powers of $u$ in (4.33), we obtain:

\begin{align*}
C^j_0 &= d_0 \\  
C^j_i &= d_i - \lambda_j C^j_{i-1} \\  
C^j_{n-1} &= d_{n-1}/\lambda_j
\end{align*}

The connection between (4.30) and the elementary symmetric functions is obvious once we realize that the coefficients of the polynomial in (4.30) are the symmetric functions of the $n$ eigenvalues. Thus,

\begin{equation}
Q(n) = f_0(\lambda_1, \ldots, \lambda_n) + f_1(\lambda_1, \ldots, \lambda_n) u + \ldots + f_n(\lambda_1, \ldots, \lambda_n) u^n
\end{equation}

Hence, the equations (4.34-4.36) become:

\begin{align*}
C^j_0 &= f_0(\lambda_1, \ldots, \lambda_n) = 1 \\  
C^j_i &= f_i(\lambda_1, \ldots, \lambda_n) - \lambda_j C^j_{i-1} \\  
C^j_{n-1} &= f_n(\lambda_1, \ldots, \lambda_n) / \lambda_j
\end{align*}

We note here that in the computation of the elementary symmetric functions, we need not compute $f_n(\lambda_1, \ldots, \lambda_n)$. This is because (4.39) is also valid for $i=n-1$ as can be
seen from (4.33). Specifically, we have

\[ C_{n-1}^j = f_{n-1}(\lambda_1, \ldots, \lambda_n) - \lambda_j C_{n-2}^j \]  \hspace{1cm} (4.41)

and thus \( f_n(\lambda_1, \ldots, \lambda_n) \) is not needed at all. An interesting consequence of (4.41) and (4.40) is that

\[ f_n = \lambda_j (f_{n-1} - \lambda_j f_{n-2} + \lambda_j^2 f_{n-3} - \cdots + (-1)^{n-1} \lambda_j^{n-1}) \]  \hspace{1cm} (4.42)

where we have dropped the \( \lambda \)'s in the symmetric functions.

Thus, if there is a \( j \) such that \( \lambda_j = 1 \) or a power of 2, (4.42) implies that \( f_n \) can be obtained without any multiplications.

Since our objective is to compute

\[ \bigg( \prod_{i \neq j} (1 + u \lambda_i) \bigg)^2 = \bigg( \prod_{i \neq j} (1 + u \lambda_i) \bigg)^2 = Q_j^2(n) \]  \hspace{1cm} (4.43)

to obtain the coefficients of the polynomial \( Q_j^2(n) \). Recognizing that if we represent the coefficients of \( Q_j(n) \) as a sequence of numbers \( \{C_i^j\} \), then the coefficients of \( Q_j^2(n) \) expressed as a sequence \( \{E_i^j\} \) are simply the sequence obtained from linearly convolving \( \{C_i^j\} \) with itself. If we denote linear convolution by *, then,

\[ \{E_i^j\} = \{C_i^j\} * \{C_i^j\} \]  \hspace{1cm} (4.44)

Thus we can use a Fast Fourier Transform (FFT) algorithm.
to compute \( \{E_i^j\} \). We note that since \( \{C_i^j\} \) is a sequence of length \( n \), therefore \( \{E_i^j\} \) is of length \( 2n-1 \). The FFT we want to use should preferably have a length that is equal to \( 2n-1 \). Hence, in addition to radix 2 FFT's, we may want to use prime number FFT's [62].

To summarize, the steps in obtaining the polynomial (4.29) are:

1) Compute the \( n-1 \) symmetric functions \( f_1, \ldots, f_{n-1} \)
of the eigenvalues \( \lambda_1, \ldots, \lambda_n \).

2) Use (4.38), (4.39), and (4.41) to obtain \( \{C_i^j\} \) for \( j=1,2,\ldots,n \).

3) Use FFT to obtain \( \{E_i^j\} \) for \( j=1,2,\ldots,2n-1 \).

4) For a fixed value of \( Z \), multiply each \( \{E_i^j\} \) by \( (Z-\lambda_j)\alpha_j^2 \) and sum over all \( j \) the sequence members with the same subscript to form the coefficients of the final polynomial, \( P(u) \) in (4.29), which is of degree \( 2n-1 \).

For each real root of \( P(u) = 0 \), the use of (4.16) and (4.20) gives the two Lagrange multipliers which, together with the selected value of \( Z \), can then be substituted in (4.18) to obtain a value for the Bhattacharyya distance. This process is repeated for different values of \( Z \) over the mesh in \( (\lambda_1, \lambda_n) \) and the set of \( Z, u \) that achieves the largest value of \( D_B \) is then used to obtain
the optimal transformation $t^*$. We note here that

$$u_1^2 = \frac{1}{4} \left( \sum_{1}^{a_1} \frac{a_1}{(1+u_{1i})^2} \right) \quad (4.45)$$

which is what we need in (4.18). Also, since

$$t = C(A^T(u_1I + u_2D)^{-1}/2) L^{-1} \quad (4.46)$$

where $C$ is a normalizing constant, and that $u_2 = u_1 \cdot u$, we can choose the positive value in (4.45) to be $u_1$.

We emphasize here that the algorithm presented in this section is amenable to parallel computation. Parallelism exists in the computation of the symmetric functions, in the array implementation of FFT processors, and in the evaluation of the polynomial $P(u, z)$ over several values of $z$. In the following section we present a numerical example using the algorithm developed in this section. The program was written in Fortran and run on IBM 370 computer. The FFT algorithm used is a prime number algorithm written by Kolba [63].

4.3 A Numerical Example

Let us consider a problem with the following statistics:
\[
R_1 = \begin{bmatrix}
12 & 1 & \frac{-1}{2} & 1 & 1 \\
1 & 14 & 1 & \frac{-1}{2} & 1 \\
-1 & 1 & 16 & -1 & 1 \\
2 & -1 & -1 & 12 & -1 \\
1 & 1 & 1 & -1 & 11 \\
\end{bmatrix}
\]

\[R_2 = \begin{bmatrix}
10 & 2 & 3 & 1 & 1 \\
2 & 12 & 1 & 2 & 1 \\
3 & 1 & 11 & 1 & -1 \\
1 & 2 & 1 & 9 & 1 \\
1 & 1 & -1 & 1 & 15 \\
\end{bmatrix}\]

\[
m = [1, 0, -3, 5, 2]^T\]

The minimum and maximum eigenvalues of the generalized characteristic equation (3.3) are:

\[
\lambda_1 = 0.43 \quad (4.50)
\]

\[
\lambda_n = 1.49 \quad (4.51)
\]

A mesh of 10 equally spaced points is chosen over the interval \((0.43, 1.49)\). The maximum value of the Bhattacharyya distance is found to be \(5.7\), which implies for the equal priors case a

\[
P_e \leq 0.00167 \quad (4.52)
\]

The normalized optimal transformation \(t^*\) is found to be:

\[
t^* = [0.00896, -0.0329, -0.295, 0.916, 0.27]^T \quad (4.53)
\]
5. **Sequentially Received Data**

In this section we consider situations in which the \( n \) dimensional data vector \( x \) is being received at each discrete time instant. To introduce the time dependency, the data vector \( x \) will be denoted by \( x(i) \). The linear transformations developed in Sections 3 and 4, are, of course, still applicable at a given point in time. If we denote the initial data vector by \( x(0) \), then we can, at time \( i \), apply the optimal transformations to the cumulative data \( X(i) \) which is a vector of length \((i+1)n\). Since the optimal transformations involve the computation of the inverses of the covariance matrices and the generalized characteristic eigenvalues and eigenvectors, the computational complexity measured in terms of the number of multiplications is on the order of \((i+1)^3n^3\). This then is the motivation for obtaining simpler procedures that take computational complexity into account.

It is important to point out at the outset that the data vectors \( x(i) \) we are considering in this section have statistics that are correlated in time. For, if the data vectors \( x(i) \) are such that they have the same statistics, then there is no point in considering sequential procedures except for system constraints such as memory size, channel capacity, time limitation, etc.
Because of the notational complexity introduced by the time element, we will attempt to use lower case letters to denote the instantaneous quantity and the capital letters to denote the cumulative quantity. Before we proceed further, a list of the notations for this and the subsequent sections will be given first.

5.1 Notations and Definitions

The n dimensional data vector $x(i)$ denotes the instantaneous data received at time $i$, and the $(i+1)n$ dimensional data vector $X(i)$ denotes the accumulated data at time $i$. We assume that the initial data time is $i=0$, and that the instantaneous data occupies the top slot in the vector $X(i)$. That is,

$$X(i) = (x(i), x(i-1), \ldots, x(0))^T$$

(5.1)

We assume as before that the data $X(i)$ is known to come from one of two distributions $H_1$ and $H_2$ that are Gaussian. Let the following notations denote the statistics of $x(i)$ and $X(i)$ and their linear transformations:

$$M(i) = E(X(i)/H_1 - X(i)/H_2).$$

$$m(i) = E(x(i)/H_1 - x(i)/H_2)$$

$$R_j(i) = \text{Covariance matrix of } X(i) \text{ under } H_j; \ j=1,2.$$
\[ r_j = \text{Covariance matrix of } x(i) \text{ under } H_j; \ j=1,2. \]

\[ c_j = \text{Cross covariance matrix between } X(i-l) \text{ and } x(i) \text{ under } H_j; \ j=1,2. \]

\[ T(i) = \text{Linear transformation for } X(i). \]

\[ t(i) = \text{Linear transformation for } x(i). \]

A superscript \(T\) will, as before, denote the transpose. No confusion between \(T(i)\) and the superscript \(T\) should result.

Using these notations, the covariance matrix \(R_j(i)\) has components \(R_j(i-l), c_j(i),\) and \(r_j(i)\) partitioned as follows:

\[
R_j(i) = \begin{bmatrix}
  r_j(i) & c_j^T(i) \\
  c_j(i) & R_j(i-l)
\end{bmatrix}
\]

(5.2)

When a square matrix, \(R\), is partitioned into smaller \(n \times n\) square matrices, the partitions will be denoted by \(P_{ij}(R)\) where \(ij\) is the location of the \(ij\)th partition.

For example, a \(2 \times 2\) matrix, \(R\), after partitioning looks like:

\[
R = \begin{bmatrix}
  P_{11}(R) & P_{12}(R) \\
  P_{21}(R) & P_{22}(R)
\end{bmatrix}
\]

(5.2)
Definition 5.1

A Single-step Linear Transformation (SLT) forms $T(i)$ by concatenating $t(i)$'s each of which is computed solely from the statistics of $x(i)$.

Remark 5.1

For the SLT $T^*(i) = (t^*(0), t^*(1), \ldots, t^*(i))$, where the superscript * denotes optimality with respect to a certain criterion. In this section, that criterion will be the maximization of the Bhattacharyya distance.

Definition 5.2

An Invariant Linear Transformation (ILT) is one in which $T(i) = (t(0), \ldots, t(i))$ where $t(0), t(1), \ldots, t(i)$ are all constrained to be equal.

Definition 5.3

A Two-step Linear Transformation (TLT) is one in which $T(i) = (t^*(i), T^*(i-1))$ $i \geq 1$, $T(0) = t(0)$. That is, the previous optimal transformation, $T^*(i-1)$, is retained and only $t(i)$ needs to be optimized.

We will now develop the optimal transformations for each of the above three types. The equal covariance case
and the equal mean case will be considered. Unless otherwise stated, the transformation $T(i)$ will be a row vector.

5.2 Equal Covariance Case: $R_1(i) = R_2(i) = R(i)$

Using our new notation, the Bhattacharyya distance is given by:

$$D_B(T(i)) = \frac{1}{2} \frac{(T(i)M(i))^2}{T(i)R(i)T^T(i)}$$  \hspace{1cm} (5.4)

Our problem is to maximize $D_B(T(i))$ with respect to $T(i)$ such that:

$$T(i)T^T(i) = i + 1$$  \hspace{1cm} (5.5)

A. SLT

By definition, the optimal transformation is:

$$T^*(i) = (t^*(0), t^*(1), \ldots, t^*(i))$$ where

$$t^*(j) = m^T(j)r^{-1}(j)/ \|m^T(j)r^{-1}(j)\| \hspace{1cm} j=0,1,\ldots,i$$  \hspace{1cm} (5.6)

B. ILT

The ILT that maximizes (5.2) subject to (5.3) is given by $T^*(i) = (t^*(0), t^*(1), \ldots, t^*(i))$ with
\[
\begin{align*}
&\text{To see this we simply use the definition for ILT, and observe that for } t(0) = t(1) = \ldots = t(i) = t, \ (5.4) \text{ becomes:} \\
&D_B(T(i)) = \frac{1}{8}(t \sum_{j=0}^{i} m(j))^2 / (t \sum_{q,s=0}^{qs} (R(i))^t T) \\
&\text{which has the same form as } (5.4).
\end{align*}
\]

C. TLT

The TLT that maximizes (5.4) subject to (5.5) is

\[
T^*(i) = (t^*(i), T^*(i-l)) \text{ where}
\]

\[
t^*(i) = K_1 T^*(i-l)(AT^*(i-l)m^T(i) - BT^*(i-l)c(i))r^{-1}(i) \\
\]

where

\[
A = R(i-l) - c(i)r^{-1}(i)c^T(i) \\
B = M(i-l) - c(i)r^{-1}(i)m(i)
\]

and \(K_1\) is a normalization constant such that

\[
t^*(i)t^*(i) = 1.
\]

The derivation of this result is given in Appendix 1.
5.3 Equal Mean Case: \( M(i) = 0 \)

For this case the Bhattacharyya distance reduces to:

\[
D_B(T(i)) = \frac{1}{2} \ln \frac{1}{2} \left( \frac{T(i)R_1(i)T^T(i)}{T(i)R_2(i)T^T(i)} + \frac{T(i)R_2(i)T^T(i)}{T(i)R_1(i)T^T(i)} + 2 \right)
\]

(5.13)

Our problem is again to maximize (5.13) over \( T(i) \) such that (5.5) is satisfied.

A. \textbf{SIT}

Again, by definition the optimal transformation

\[
t^*(i) = (t^*(0), t^*(1), \ldots, t^*(i))
\]

is such that:

\[
t^*(j) = t_1(j) \quad j = 0, 1, \ldots, i
\]

(5.14)

where \( t_1(j) \) is the \( t_1 \) eigenvector of the generalized characteristic equation with \( r_1(j) \) and \( r_2(j) \) as the covariance matrices.

B. \textbf{ILT}

The ILT that maximizes (5.13) subject to (5.5) is

\[
t^*(i) = (t^*(0), t^*(1), \ldots, t^*(i)) \text{ with }
\]

\[
t^*(0) = t^*(1) = \ldots = t^*(i) = t^* = t_1(\Sigma)
\]

(5.15)

where \( t_1(\Sigma) \) denotes the \( t_1 \) eigenvector of the generalized characteristic equation with \( \sum_{q,s=0}^{n} P_{qs}(R_1(i)) \).
and \( \sum_{q,s=0}^{n} P_{qs} (R_{2}(i)) \) as the covariance matrices.

To see this we simply note that:

\[
T(i)R_j(i)T^T(i) = t( \sum_{q,s=0}^{n} P_{qs} (R_{j}(i)) ) t^T
\]

(5.16)

which when substituted in (5.13) gives (5.15).

C. TLT

The TLT for the equal mean case does not have a closed form solution. Since the maximization of (5.13) depends on the ratio of the quadratic terms, let:

\[
f(T(i)) = T(i)R_1(i)T^T(i)/T(i)R_2(i)T^T(i)
\]

(5.17)

For the TLT case (5.17) can be transformed into the following form (see Appendix 2):

\[
f(z) = \frac{zDz^T + zB + a_1}{a_2 + zz^T}
\]

(5.18)

where \( D \) is a diagonal matrix, \( B \) is a vector, and \( a_1, a_2 \) are scalars with \( a_2 > 0 \). A necessary condition for the extrema of (5.18) to satisfy is that

\[
\text{Grad } f(z) = 0
\]

(5.19)

The stationary points from (5.19) are given by:

\[
z = -\frac{B^T}{2} (D - f(z)I)^{-1}
\]

(5.20)
which when substituted back into (5.18) implies that:

\[ f(z) = \frac{n}{a_2} \left( \sum_{i=1}^{n} \frac{b_i^2}{4(f(z)-\lambda_i^2)} + a_1 \right) \]

where the \( b_i \)'s are the elements of \( B \) and \( \lambda_i \)'s are the diagonal elements of \( D \).

**Lemma 5.1**

For distinct values of \( \lambda_i \) \( i=1,2,\ldots,n \) with \( \lambda_1 < \lambda_2 < \ldots < \lambda_n \) and nonzero \( b_i \)'s, there are \( n+1 \) values of \( f(z) \), \( f_0, f_1, \ldots, f_n \) that satisfy (5.21) and that \( 0 < f_0 < \lambda_1 < f_1 < \ldots < \lambda_n < f_n \).

**Proof:**

Let \( g(f(z)) = f(z)a_2 - a_1 - \sum_{i=1}^{n} \frac{b_i^2}{4(f(z)-\lambda_i^2)} \)

then

\[ \frac{dg(f(z))}{df(z)} = a_2 + \sum_{i=1}^{n} \frac{b_i^2}{4(f(z)-\lambda_i^2)^2} \]

which is strictly monotonic for \( \lambda_i < f(z) < \lambda_i+1 \) \( i = 1,2,\ldots,n \) where \( \lambda_{n+1} \) is taken to be + . Since \( g(f(z)) \) is continuous for \( f(z) \) inside a given interval \( (\lambda_i,\lambda_{i+1}) \), and that since \( g(f(z)) \) necessarily has to cross the f axis at some point inside each interval, we conclude that there are \( n+1 \) such points and that \( f_0 < \lambda_1 < f_1 < \ldots < \lambda_n < f_n \). The condition \( f_0 > 0 \) follows from the fact that the covariances matrices are
assumed to be positive definite.

Thus, the TLT, expressed in terms of z, that maximizes (5.18) is:

\[ z^* = \frac{-B^T}{2} (D - f^0 I)^{-1} \]  \hspace{1cm} (5.24)

where

\[ f^0 = \max \left( f_0 + 1/f_0, \frac{f_n}{f_n} + 1/f_n \right) \]  \hspace{1cm} (5.25)

The above lemma can also be seen easily by graphing the functions

\[ g_1(f(z)) = f(z) \]  \hspace{1cm} (5.26)

\[ g_2(f(z)) = \left( \sum_{i=1}^{n} \frac{b_i^2}{4(f(z) - \lambda_i^2)} + a_1 \right) / a_2 \]  \hspace{1cm} (5.27)

Approximate solutions for special cases can be obtained. For example, if \( f_0 < \lambda_1 \), then

\[ g(f_0) \approx \left( a_1 - \sum_{i=1}^{n} \frac{b_i^2}{4\lambda_i} \right) / a_2 \]  \hspace{1cm} (5.28)

which we must check to make sure that it is > 0.

The SIT, IIT, and TTT described in this section can be viewed from two angles. If at the time instant \( i \) we employ \( T^*(i) \), then a \((i+1)n\) dimensional data \( X(i) \) is reduced to a scalar. If, however, we use the \( t^*(i) \) for each of the \( i+1 \) data vectors \( x(j) j=0,1,...,i \), then we
would obtain an \( n \) to 1 data reduction for each \( x \). These three types of transformations differ in how the past history of the data is utilized. The SLT, for example, ignores all the past data in arriving at the transformations.

In the following section we will consider data vectors that can be modeled by Gauss-Markov processes. An attempt will be made to link the Bhattacharyya distance with the Kalman filtering equations.
6. **Gauss-Markov Model**

Let the n-dimensional vector observations \( y(\cdot) \) of a signal process \( x(\cdot) \) be described by:

\[
y(i) = C(i)x(i) + D(i)v(i) \quad i=0,1,2... \quad (6.1)
\]

where the random process \( v(\cdot) \) is uncorrelated with the signal process. The matrices \( C(i) \) and \( D(i) \) are known, and we assume that \( v(\cdot) \) is a zero mean white Gaussian process with covariance,

\[
E[v(i)v^T(j)] = L(i) \delta(i-j) \quad (6.2)
\]

where \( \delta \) is the Dirac delta function.

We assume that \( y(\cdot) \) is a Gaussian process described by one of two hypotheses \( H_1, H_2 \).

\[
H_1 : \quad y(\cdot) = y_1(\cdot) = N(m_1(\cdot), r_1(\cdot)) \quad (6.3)
\]

\[
H_2 : \quad y(\cdot) = y_2(\cdot) = N(m_2(\cdot), r_2(\cdot)) \quad (6.4)
\]

Let the n-dimensional signal process be described by:

\[
x(i+1) = A(i)x(i) + B(i)u(i) \quad i=0,1,2... \quad (6.5)
\]

where \( A(i), B(i) \) are known matrices, and we assume that \( u(\cdot) \) is a zero mean white Gaussian process with covariance

\[
E[u(i)u^T(j)] = Q(i) \delta (i-j) \quad (6.6)
\]
We assume that $u(\cdot)$ and $x(\cdot)$ are uncorrelated, and that $u(\cdot)$ and $v(\cdot)$ are also uncorrelated.

An objective of this section is to express the optimal transformations obtained previously in terms of the parameters of the dynamical systems. Our first step will be to show the relationship between the Bhattacharyya distance and the Kalman filtering equations. The procedure we will use follows that used by Schwepp [64].

We will denote, as in Section 5, the cumulative observations at time $i$ by $Y(i)$. Thus, $Y(i)$ is $N(M_1(0), R_1(0))$ under $H_1$ and $N(M_2(\cdot), R_2(\cdot))$ under $H_2$.

Let the covariances $R_j(i)$ be described as in (5.2), and let:

$$
\bar{Y}_j(i) = E(y_j(i)/Y_j(i-1)) \quad j=1,2
$$

(6.7)

then,

$$
\mathbf{\bar{Y}}_j(i) = m_j(i) + c_j^T(i)R_j^{-1}(i-1)(Y_j(i-1) - M_j(i-1))
$$

(6.8)

Define:

$$
S_j(i) = E[(y_j(i) - \bar{Y}_j(i))(y_j(i) - \bar{Y}_j(i))^T] \quad j=1,2
$$

(6.9)

which, because of (6.8), equals to:

$$
S_j(i) = r_j(i) - c_j^T(i)R_j^{-1}(i-1)c_j(i) \quad j=1,2
$$

(6.10)

where $R_j^{-1}(i-1)$ can be partitioned as:
\[
\begin{bmatrix}
S_{j-1}^{-(i-1)} & -S_{j-1}^{-(i-1)}c_j^T(i-1)R_j^{-1}(i-2) \\
-R_j^{-1}(i-2)c_j(i-1)S_j^{-(i-1)} & R_j^{-1}(i-2)+R_j^{-1}(i-2)c(i-1)S_j^{-(i-1)}
\end{bmatrix}
\]

\[j=1,2\]  \hspace{1cm} (6.11)

We now define:

\[R_0(i) = R_1(i) + R_2(i)\]  \hspace{1cm} (6.12)
\[r_0(i) = r_1(i) + r_2(i)\]  \hspace{1cm} (6.13)
\[c_0(i) = c_1(i) + c_2(i)\]  \hspace{1cm} (6.14)
\[y_0(i) = y_1(i) + y_2(i)\]  \hspace{1cm} (6.15)
\[Y_0(i) = Y_1(i) + Y_2(i)\]  \hspace{1cm} (6.16)

Then (6.10) also holds for \(j=0\).

It is well known that the determinant of \(R_j(i)\) can be expressed by:

\[|R_j(i)| = |R_j(i-1)| \cdot |S_j(i)|\]  \hspace{1cm} (6.17)

Letting the difference in the means be denoted by:

\[M(i) = M_1(i) - M_2(i)\]  \hspace{1cm} (6.18)
\[m(i) = m_1(i) - m_2(i)\]  \hspace{1cm} (6.19)

then, using (6.11) for \(j=0\), we can write:

\[M^T(i)R_0^{-1}(i)M(i) = M^T(i-1)R_0^{-1}(i-1)M(i-1) + H_0(i)S_0^{-1}(i)H_0^T(i)\]  \hspace{1cm} (6.24)
Since the Bhattacharyya distance for Gaussian hypotheses is given by:

\[ D_B(i) = \frac{1}{4}(M^T(i)R_0^{-1}(i)M(i) + \ln R_0(i) R_1(i) R_2(i)) \]  

(6.22)

the use of (6.21) and (6.17) imply that:

\[ D_B(i) - D_B(i-1) = \frac{1}{4}(H_0(i)S_0^{-1}(i)H_0^T(i) + \ln S_0(i) S_1(i) S_2(i)) \]  

(6.23)

where

\[ D_B(-1) = 0 \]  

(6.24)

To link (6.23) with the parameters of the Gauss-Markov model, we need the discrete Kalman filtering equations.

Suppose that we want to find an estimate \( \bar{x}(i+1) \) of \( x(i+1) \) from the accumulated observations \( Y(i) \) such that the mean squared error:

\[ E[(x(i+1) - \bar{x}(i+1))^T(x(i+1) - \bar{x}(i+1))] \]  

(6.25)

is minimized. By using the orthogonality principle, the estimator \( \bar{x}(i+1) \) can be expressed iteratively as:

\[ \bar{x}(i+1) = (A(i) - F(i)C(i)) \bar{x}(i) + F(i)y(i) \]  

(6.26)

where,

\[ F(i) = A(i)P(i)C^T(i)[C(i)P(i)C^T(i) + D(i)L(i)D^T(i)]^{-1} \]  

(6.27)
and:
\[ P(i+1) = E[(x(i+1) - \bar{x}(i+1))(x(i+1) - \bar{x}(i+1))^T] \]
\[ = (A(i) - F(i)C(i))P(i)(A(i) - F(i)C(i))^T + \\
B(i)Q(i)B^T(i) + F(i)D(i)L(i)D^T(i)F^T(i) \quad (6.28) \]
is the error covariance matrix.
To solve (6.26-6.28), the initial conditions \( \bar{x}(0), P(0) \)
should also be specified.

It is well known that the estimator \( \bar{x}(i+1) \) is also expressed by:
\[ \bar{x}(i+1) = E \left( x(i+1)/Y(i) \right) \quad (6.29) \]
which, using (6.1), implies that,
\[ \bar{y}(i) = C(i) \bar{x}(i) \quad (6.30) \]
Substituting (6.1) and (6.30) in (6.9) gives:
\[ S_j(i) = C_j(i)P_j(i)C_j^T(i) + D(i)L(i)D^T(i) \quad j=1,2 \quad (6.31) \]
To obtain a similar expression for \( S_0(i) \), we need to define a hypothetical system:
\[ x_0(i+1) = A_0(i)x_0(i) + B_0(i)u_0(i) \quad (6.32) \]
\[ y_0(i) = C_0^T(i)x_0(i) + 2D(i)v(i) \quad (6.33) \]
where:
\[ x_0(i) = (x_1^T(i), x_2^T(i))^T \quad (6.34) \]
\[ u_0(i) = (u_1^T(i), u_2^T(i))^T \quad (6.35) \]
\[ A_0(i) = \begin{bmatrix} A_1(i) & 0 \\ 0 & A_2(i) \end{bmatrix} \quad (6.36) \]
\[ B_0(i) = \begin{bmatrix} B_1(i) & 0 \\ 0 & B_2(i) \end{bmatrix} \quad (6.37) \]
\[ Q_0(i) = \begin{bmatrix} Q_1(i) & 0 \\ 0 & Q_2(i) \end{bmatrix} \quad (6.38) \]
\[ C_0(i) = [C(i), C(i)]^T \quad (6.39) \]

The Kalman filtering equations for this system are given below:

\[ \bar{x}_0(i+1) = (A_0(i) - F_0(i)C_0^T(i))\bar{x}_0(i) + F_0(i)y_0(i) \quad (6.40) \]
\[ F_0(i) = A_0(i)P_0(i)C_0(i)[C_0^T(i)P_0(i)C_0(i) + 4D(i)L(i)D^T(i)]^{-1} \quad (6.41) \]
\[ P_0(i+1) = (A_0(i) - F_0(i)C_0^T(i))P_0(i)(A_0(i) - F_0(i)C_0^T(i))^T + B_0(i)Q_0(i)B_0^T(i) + 4F_0(i)D(i)L(i)D^T(i)F_0^T(i) \quad (6.42) \]

Thus,

\[ S_0(i) = C_0^T(i)P_0(i)C_0(i) + 4D(i)L(i)D^T(i) \quad (6.43) \]
Equations (6.31) and (6.43) provide the link between the dynamical system parameters and the difference between two Bhattacharyya distances, called the incremental Bhattacharyya distance, expressed by (6.23). The computations of (6.31) and (6.43) are iterative. However, to compute (6.23) iteratively, the vector $H_0$ also needs to have an iterative expression. We will concentrate on the last term in (6.20). Appendix 3 shows that if the matrices $A_1(i) = A(i)$ for all $i$, and that the inverse of the matrix $C$ exists, then:

$$M^T(i)R_0^{-1}(i)C_0(i+1) = m^T(i)f^T(i) + M^T(i-1)R_0^{-1}(i-1) \cdot$$

$$C^T_0(i)(C(i)A(i)C^{-1}(i) - f(i))^T \quad (6.45)$$

where

$$f(i) = C(i)A(i)C^{-1}(i)(I - 2D(i)L(i)D^T(i)S_0^{-1}(i)) \quad (6.45)$$

Thus, at this stage, the quantities $S_0(i), S_1(i), S_2(i), \text{ and } H_0(i)$ can all be computed iteratively. This means, of course, that (6.23) can also be computed in an iterative manner.

6.1 **Linear Transformation**

The incremental Bhattacharyya distance in the transformed space can not be expressed as in (6.23). Since we
already know the form of the optimal n to l linear transformations, let us find the relationships between \(t^*(i)\) and \(t^*(i+1)\).

A. **Equal Covariance Case**

Let us consider

\[
t^*(i) = M^T(i)R^{-1}(i) \quad (6.46)
\]

\[
t^*(i+1) = M^T(i+1)R^{-1}(i+1) \quad (6.47)
\]

where the norming constants have been set to 1. Using (6.11), we see that (6.47) can be expressed as:

\[
t^*(i+1) = (m^T(i+1)S^{-1}(i+1) - M^T(i)R^{-1}(i)c(i+1)S^{-1}(i+1),

- m^T(i+1)S^{-1}(i+1)c^T(i+1)R^{-1}(i) + M^T(i)R^{-1}(i)

+ M^T(i)R^{-1}(i)c(i+1)S^{-1}(i+1)c^T(i+1)R^{-1}(i))
\]

(6.48)

Substituting (6.45) into (6.47), we obtain:

\[
t^*(i+1) = (d(i+1), t^*(i) - d(i+1)s(i+1)) \quad (6.49)
\]

where,

\[
d(i+1) = m^T(i+1)S^{-1}(i+1) - t^*(i)c(i+1)s^{-1}(i+1) \quad (6.50)
\]

\[
s(i+1) = c^T(i+1)R^{-1}(i) \quad (6.51)
\]

and \(S(i)\) is given by (6.10).
Thus, in effect, we are using (6.11) for computing the inverse of \( R(i+1) \). The matrix \( S(i+1) \) can be computed from (6.10) or form (6.31) when the system is Gauss-Markov. The advantages of (6.49-6.51) are firstly, that the inversion of \( R(i+1) \) is reduced to inverting an \( n \times n \) matrix \( S(*) \), and secondly, that when (6.31) is used, the computation of the error covariance matrix can be done off-line. Also, when \( P(*) \) reaches a steady state value, the computation of \( S(*) \) is further simplified.

The incremental Bhattacharyya distance for the equal covariance case can now be expressed by:

\[
D_{B,t^*}(i) - D_{B,t^*}(i-1) = \frac{1}{8} \left( (H_0(i)S^{-1}(i)H_0^T(i))^2 + 2(H_0(i)S^{-1}(i)H_0^T(i))t^*(i-1)M(i-1) \right) / (t^*(i)R(i-1)t^*T(i-1)) \tag{6.52}
\]

where \( H_0(i) \) can be computed iteratively if (6.44) and (6.45) are used in (6.20). Equation (6.52) may be used to select the data size for transformation by noting the point in time when (6.52) becomes small.

B. **Equal Mean Case**

The basic difficulty in this case is that in general, there is no apparent relationship existing between the
eigenvalues of a matrix $R(i)$ and those of $R(i+1)$. The same comment applies to the eigenvectors. A possible alternative is to first obtain an estimate of the dimensionality of the transformation vector by comparing the values of the Bhattacharyya distance of the estimator $\bar{x}(i)$ as $i$ increases. Once the dimensionality is fixed, then the method in Section 4 can be used to obtain the optimal transformation.
7. **Stationary Gaussian Processes**

In this section we will consider the situation in which the data are produced from one of two zero mean stationary Gaussian random processes. If we let $R_1^n$ and $R_2^n$ be the $n \times n$ covariance matrices under hypotheses $H_1$ and $H_2$ respectively, then the Bhattacharyya distance for the equal mean case is expressed by:

$$2D_B(n) = \ln \left| \frac{1}{2} R_1^n + R_2^n \right| - \frac{1}{2} \ln | R_1^n | - \frac{1}{2} \ln | R_2^n | \quad (7.1)$$

Our objective in this section is to seek an expression for $D_B(n)$ as the number of data, $n$, becomes very large. Let $D_B$ denote the asymptotic Bhattacharyya distance, then

$$D_B = \lim_{n \to \infty} n^{-1} D_B(n) \quad (7.2)$$

The significance of (7.2) is that the probability of error can then be bounded by:

$$P_e \leq 1/2 \exp \left[ -n(D_B + O(1)) \right] \quad (7.3)$$

where $O(1) \to 0$ as $n \to \infty$.

Assume that $R_J^n$, $J=1,2$, have elements $r_J(i | i-k |)$ where $i$ and $k$ are the row and column indices. The spectral densities of $R_J^n$ are given by:

$$f_J(\theta) = \sum_{k=-\infty}^{\infty} r_J(k) \exp (jk \theta) \quad j = 1, 2 \quad (7.4)$$
where \( f_J(\theta) \) and \( \theta \) are assumed to be such that:

\[
0 < u_1 < f_J(\theta) < u_2 < \infty \tag{7.5}
\]

\[
\theta \in [0, 2\pi] \tag{7.6}
\]

Gray [65] has shown that if \( R^n_j \) has eigenvalues \([\lambda_j, k = 1, 2, \ldots, n]\), then (7.1) and (7.2) imply that:

\[
\lim_{n \to \infty} n^{-1} \sum_{k=1}^{n} F(\lambda_j, k) = (2\pi)^{-1} \int_0^{2\pi} F(f_J(\theta))d\theta \tag{7.7}
\]

where \( F(x) \) is a continuous function on \( x \in [u_1, u_2] \). The above expression (7.7) gives the asymptotic distribution of eigenvalues of Toeplitz matrices. To use (7.7) for our purpose, we note that:

\[
\ln |R^n_j| = \sum_{k=1}^{n} \ln \lambda_j, k \quad J=1, 2 \tag{7.8}
\]

Since the matrices \( R^n_1 \) and \( R^n_2 \) are assumed to be positive definite, the conditions on \( F(x) \) in (7.7) are satisfied. Hence, we can apply (7.7) to (7.8) to obtain:

\[
\lim_{n \to \infty} n^{-1} \ln |R^n_j| = \lim_{n \to \infty} n^{-1} \sum_{k=1}^{n} \ln \lambda_j, k \int_0^{2\pi} F(f_J(\theta))d\theta \quad J=1, 2 \tag{7.9}
\]

If we let

\[
R^n_0 = R^n_1 + R^n_2 \tag{7.10}
\]

then the spectral density of \( R^n_0 \), denoted by \( f_0(\theta) \), is
given by:

\[ f_0(\theta) = f_1(\theta) + f_2(\theta) \quad (7.11) \]

Suppose the eigenvalues of \( \frac{1}{2} R_n^0 \) are \( \{u_k : k=1,2,\ldots,n\} \), then applying \((7.7)\) to \( \ln \frac{1}{2} R_n^0 \) gives:

\[
\lim_{n \to \infty} n^{-1} \ln |\frac{1}{2} R_n^0| = \lim_{n \to \infty} n^{-1} \sum_{k=1}^{n} \ln u_k = (2\pi)^{-1} \int_{0}^{2\pi} \ln(f_1(\theta) + f_2(\theta)) d\theta \quad (7.12)
\]

Thus, using \((7.9)\) and \((7.12)\) in \((7.1)\), we can express the asymptotic Bhattacharyya distance \( D_B \) by:

\[
D_B = (4\pi)^{-1} \int_{0}^{2\pi} \left[ \ln(f_1(\theta) + f_2(\theta)) - \frac{1}{2} \ln f_1(\theta) - \frac{1}{2} \ln f_2(\theta) \right] d\theta \quad (7.13)
\]

Let us denote \( x_p^J, J=1,2 \), as the \( J \) th sample of an \( n \)-dimensional stationary Gaussian process, distributed according to one of two hypotheses \( H_1, H_2 \). If we let \( r_{pq}(|i-k|) \), where \( i, k \) are indices, denote the cross-correlations of the \( p \) th and \( q \) th samples under hypothesis \( J \), then the corresponding spectral densities are:

\[
f_{pq}^J(\theta) = \sum_{m=-\infty}^{\infty} r_{pq}^J(m) \exp(j\theta m) \quad J=1,2 \quad (7.14)
\]

\[ p, q = 1, 2, \ldots, n \]

Let \( F_J(\theta) \) be the \( n \times n \) matrix with \( \{f_{pq}^J(\theta), \ p, q=1, 2, \ldots, n\} \) as the elements.

The vector process \( \{x_p, p=1,2,\ldots,n\} \) is reduced to
a scalar process \( \{Z_p \mid p=1,2,\ldots,n\} \) by a linear transformation, \( t \),:

\[
Z_p = t x_p \tag{7.15}
\]

where \( t \) is a \( 1 \times n \) row vector. The process \( \{Z_p\} \) is still Gaussian with spectral density:

\[
f_J(\theta) = t F_J(\theta) t^T \quad J=1,2 \tag{7.16}
\]

The transformed Bhattacharyya distance can now be expressed in terms of the transformation, \( t \), as:

\[
D_{B,t} = (4\pi)^{-1} \int_0^{2\pi} \left[ \ln(\frac{1}{2} t(F_1(\theta) + F_2(\theta) t^T) - \frac{1}{2} \ln t F_1(\theta) t^T - \frac{1}{2} \ln t F_2(\theta) t^T \right] d\theta \tag{7.17}
\]

We will not attempt to obtain the optimal transformation \( t^* \), since to do so we need the specific spectral densities, \( F_1(\theta) \) and \( F_2(\theta) \).
8. **Conclusions and Suggestions for Further Research**

The feature selection problem for statistical data was studied in Part II of this thesis. We have concentrated on the use of distance measures as optimality criteria for obtaining linear transformations.

In Section 2, we presented several distances, some of which have built in penalty functions or distortion measures. In Section 3, these distances were applied to Gaussian distributed data. In Section 4, the Bhattacharyya distance was examined in detail, and an algorithm for reducing a data vector to a scalar was presented. The algorithm was also noted to have parallel computational structures. In Section 5, we discussed sequential feature selection procedures when constraints on the linear transformations are imposed. In Section 6, we considered the advantages offered when the data vectors are generated by Gauss-Markov processes. In Section 7, the use of spectral densities, when the data were generated by stationary Gaussian processes in the Bhattacharyya distance, was described.

The selection of a particular distance measure for data reduction depends on the distance's characteristics such as calculability, bounds on the probability of
error, and sensitivity to the underlying distributions. From Sections 6 and 7, we see that the nature of the data can also be explored to offer certain computational advantages. The feasibility of constructing an actual hardware to perform the algorithm described in Section 4, is an interesting project that needs to be explored. The use of Kalman filtering equations and the spectral density functions in computing the optimal transformation also merits further considerations. The use of the incremental Bhattacharyya distance for data size selection, as mentioned in Section 6, also needs to be explored for practical situations. Finally, distances with penalty functions offer interesting theoretical considerations. Their potential as feature selection performance criteria requires a fuller treatment.
REFERENCES


Theory 12, 206-214.


Appendix 1

We want to show that the $t^*(i)$ in (5.9) of Part II maximizes (5.4) subject to (5.12).

Let the mean vector $M(i)$ and the covariance matrix $R(i)$ be partitioned such that,

$$M(i) = \begin{bmatrix} m(i) \\ M(i-1) \end{bmatrix} \quad (A.1.1)$$

$$R(i) = \begin{bmatrix} r(i) & c^T(i) \\ c(i) & R(i-1) \end{bmatrix} \quad (A.1.2)$$

Let

$$a = T*(i-1)M(i-1) \quad (A.1.3)$$

$$b = c^T(i)T^T*(i-1) \quad (A.1.4)$$

$$c = T*(i-1)R(i-1)T^T*(i-1) \quad (A.1.5)$$

then for $T(i) = (t(i), T^*(i-1)$ the Bhattacharyya distance (5.4) becomes:

$$D_B(T(i)) = 1/8 \ (a+t(i)m(i))^2/(t(i)r(i)t^T(i)+2t(i)b+c) \quad (A.1.6)$$

If we denote the Cholesky decomposition of $r(i)$ by
\[ r(i) = w(i)w^T(i) \quad \text{(A.1.7)} \]

then the transformations
\[ d(i) = w^{-1}(i)m(i) \quad \text{(A.1.8)} \]
\[ e(i) = w^{-1}(i)b \quad \text{(A.1.9)} \]
\[ t(i) = y(i)w^{-1}(i) \quad \text{(A.1.10)} \]

changes (A.1.6) into:
\[
D_B(y(i)) = \frac{1}{8} (a+y(i)d(i))^2 \sqrt{(e^T(i)e(i)) + (y(i)+e^T(i)) \\
(y(i)+e^T(i))} \quad \text{(A.1.11)}
\]

Denoting \( z(i) = y(i) + e^T(i) \)
\[
D_B(z(i)) = \frac{1}{8} (a-e^T(i)d(i)+z(i)d(i))^2 \sqrt{(c-e^T(i)e(i) + z(i)z^T(i))} \quad \text{(A.1.13)}
\]

If we fix the norm of \( z(i) \) to equal \( Z^2 \), then the maximum of \( D_B(z(i)) \) occurs when
\[ z(i) = \text{SGN}(a-e^T(i)d(i)) \frac{Zd^T(i)}{(d^T(i)d(i))^\frac{1}{2}} \quad \text{(A.1.14)} \]

Substituting (A.1.14) into (A.1.13), we find that the value of \( Z \) that maximizes (A.1.13) occurs when
\[ Z = \left( \text{SGN}(a - e^T(i)d(i)) \right) \left( c - e^T(i) \right) \left( d^T(i)d(i) \right)^{\frac{1}{2}} / \]
\[ (a - e^T(i)d(i)) \]

(A.1.15)

Thus by back transforming (A.1.15) into (A.1.12) and (A.1.10), we obtain the desired result for the TLT in Section 5.2.
Appendix 2

We want to show the transformation of (5.17) of Part II into (5.18).

Let $d_j(i) = c_j(i) T^T(i-1) \quad j=1,2 \quad (A.2.1)$

$$e_j = T^T(i-1) R_j(i-1) T^T(i-1) \quad j=1,2 \quad (A.2.2)$$

then (5.17) can be expressed by:

$$f(t(i)) = \frac{t(i)r_1(i)t^T(i)+2t(i)d_1(i)e_1}{t(i)r_2(i)t^T(i)+2t(i)d_2(i)+e_2} \quad (A.2.3)$$

If we now let $r_2 = LL^T$, $r_1 = LD L^T$, and $y(i) = t(i)L$, then (A.2.3) becomes:

$$f(y(i)) = \frac{y(i)Dy^T(i)+2y(i)L^{-1}d_1(i)+e_1}{y(i)y^T(i)+2y(i)L^{-1}d_2(i)+e_2} \quad (A.2.4)$$

Letting:

$$a_1 = (d_2^T(i)L^{-T}DL^{-1}d_2(i)) - 2d_2^T(i)L^{-T}L^{-1}d_1(i)+e_1 \quad (A.2.5)$$

$$a_2 = e_2 - d_2^T(i)L^{-T}L^{-1}d_1(i) \quad (A.2.6)$$

$$B = 2(L^{-1}d_1(i) - D L^{-1}d_2(i)) \quad (A.2.7)$$

$$z(i) = y(i) + d_2^T(i) L^{-T} \quad (A.2.8)$$

and substituting (A.2.5-A.2.8) into (5.17), gives (5.18).
Appendix 3

To show (6.43) and (6.44) of Part II, we begin by partitioning $c_0^T(i+1)$:

$$c_0^T(i+1) = [E(a(i+1)), E(b(i+1))] \quad (A.3.1)$$

where

$$a(i+1) = [y_1(i+1)-m_1(i+1), y_2(i+1)-m_2(i+1)] \cdot$$

$$[y_1(i+1)-m_1(i+1), y_2(i+1)-m_2(i+1)]^T \quad (A.3.2)$$

and

$$b(i+1) = [y_1(i+1)-m_1(i+1), y_2(i+1)-m_2(i+1)] \cdot$$

$$[Y_1(i-1)-M_1(i-1), Y_2(i-1)-M_2(i-1)]^T \quad (A.3.3)$$

Using (A.3.1) - (A.3.3), we obtain:

$$c_0^T(i+1)R_0^{-1}(i)M(i) = [E(a(i+1)) - E(b(i+1))R_0^{-1}(i-1)c_0(i)] \cdot$$

$$S_0^{-1}(i)m(i) + [E(b(i+1))R_0^{-1}(i-1)c_0(i) -$$

$$E(a(i+1))]S_0^{-1}(i)c_0^T(i)R_0^{-1}(i-1)M(i-1) +$$

$$E(b(i+1))R_0^{-1}(i-1)M(i-1) \quad (A.3.4)$$

For (A.3.4) to be iterative, the third term on the right hand side must contain the group $c_0^T(i)R_0^{-1}(i-1)M(i-1)$. This means that $E(b(i+1))$ must contain the term $c_0^T(i)$.
Let us investigate $E(b(i+1))$ by checking the system
equations which we repeat here:

$$x_j(i+1) = A_j(i)x_j(i) + B_j(i)u_j(i) \quad j=1,2 \quad (A.3.5)$$

$$y_j(i) = C(i)x_j(i) + D(i)v(i) \quad j=1,2 \quad (A.3.6)$$

The above expressions imply that:

$$y_j(i+1) = C(i)A_j(i)x_j(i)+C(i)B_j(i)u_j(i) + D(i)v(i+1) \quad (A.3.7)$$

If $C(i)$ is square and invertible, $(A.3.7)$ becomes:

$$y_j(i+1) = C(i)A_j(i)C^{-1}(i)y_j(i) - C(i)A_j(i)C^{-1}(i)D(i)v(i) +$$

$$C(i)B_j(i)u_j(i) + D(i)v(i+1) \quad (A.3.8)$$

If $C(i)$ is not square, but $(C^T(i)C(i))^{-1}$ exists, then a
similar equation to $(A.3.8)$ can be derived.

From $(A.3.8)$ we obtain:

$$E(b(i+1)) = C(i)A_1(i)C^{-1}(i)E[(y_1(i)-m_1(i))(Y_1(i-1)-\bar{m}_1(i-1))$$

$$M_1(i-1)^T] + C(i)A_2(i)C^{-1}(i)E[(y_2(i)-m_2(i))$$

$$(Y_2(i-1)-M_2(i-1)^T] \quad (A.3.9)$$

If $A_1(i) = A_2(i) = A(i)$, then $(A.3.9)$ becomes:
\[ E(b(i+1)) = C(i)A(i)C^{-1}(i)c_0^T(i) \]  \hspace{1cm} (A.3.10)

which is what we want.

Substituting (A.3.10) into (A.3.4); and expanding \( E(a(i+1)) \), we obtain the desired equations (6.43) and (6.44).