Naval Research Laboratory
Washington, DC 20375-5000

NRL Memorandum Report 6340

Optimal and Robust Memoryless Discrimination from Dependent Observations

DOUGLAS W. SAUDER

Advanced Techniques Branch
Tactical Electronic Warfare Division

November 9, 1988

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NRL Memorandum Report 6340	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Naval Research Laboratory 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Washington, DC 20375-5000	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Office of Naval Technology 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) Arlington, VA 22217	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 62113N; PROJECT NO: RE13T30; TASK NO: ; WORK UNIT ACCESSION NO: DN156-056

11. TITLE (Include Security Classification)
Optimal and Robust Memoryless Discrimination from Dependent Observations

12. PERSONAL AUTHOR(S)
Sauder, D.W.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM 5/87 TO 5/88	1988 November 9	96

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Signal detection; Hypothesis testing

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Memoryless discrimination is a method for binary hypothesis testing in which a test statistic is computed by passing each observation through a memoryless nonlinearity and summing the outputs. Under certain conditions, including a mixing condition on the observed process, such a test statistic becomes asymptotically Gaussian, thus permitting the error probabilities to be approximated. In this thesis, asymptotic performance measures are derived as functionals of the nonlinearities, and in this way memoryless discrimination is well formulated as an optimization problem. Both the Neyman-Pearson and minimax problems are considered. In all, four different performance measures are derived under different problem formulations, and in each case, the optimal nonlinearity is shown to be the solution of an integral equation. Results for minimax robustness are presented for three of the performance measures. Performance results from numerical simulations for each of the nonlinearities derived here as well as the optimal iid nonlinearity are presented and discussed.

Keywords: Signal detection. (KR) ←

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Thomas J. Calomiris	22b. TELEPHONE (Include Area Code) (202) 767-3178 22c. OFFICE SYMBOL Code 5755.2

DD Form 1473, JUN 86 — Previous editions are obsolete — SECURITY CLASSIFICATION OF THIS PAGE

S/N 0102-LF-014-6603

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acknowledgement

I would like to thank my thesis advisor Dr. Evaggelos Geraniotis for the time he spent in assisting me to complete this work. Dr. Geraniotis has been an outstanding mentor as well as a thesis advisor, and I would strongly recommend him to anyone who is looking for an advisor.

Thanks also to Jim Dunyak, Tom Calomiris, Joe Lawrence, and others at NRL for their support.

Finally, I would like to acknowledge the Naval Research Laboratory and the Office of Naval Research which provided the funding for my fellowship appointment under Grant N00014–85–G–0213, and the Systems Research Center at the University of Maryland which administered the fellowship program.

OPTIMAL AND ROBUST MEMORYLESS DISCRIMINATION FROM DEPENDENT OBSERVATIONS

CHAPTER 1

INTRODUCTION

In this thesis we will consider various special cases of the binary hypothesis testing problem, which may be informally described as follows: One observes some random event and wishes to decide on the basis of the observation between two hypotheses which concern the nature of the random event. In all cases that we will consider, the random event is modeled by a discrete-time random process $\{X_i\}$ and the two hypotheses concern the probability distribution of the process. More specifically, we consider the following two hypotheses:

$$\begin{aligned} H_0&: \ \{X_i\}_{i=1}^n \text{ has a density } f_0^{(n)}(\mathbf{x}) \\ H_1&: \ \{X_i\}_{i=1}^n \text{ has a density } f_1^{(n)}(\mathbf{x}) \end{aligned} \tag{1.1}$$

where $\mathbf{x}$ denotes the n-tuple $(x_1, \ldots, x_n)$. A decision rule for this hypothesis testing problem is a measurable mapping d which maps the observation space $\mathbf{R}^n$ into the set $\{0,1\}$, the interpretation being that if $\mathbf{x}$ is observed and $d(\mathbf{x}) = i$, then one decides that H_i is true. Such a mapping may also be referred to in this thesis as a test, a receiver, or a discriminator. If $f_1^{(n)}$ is absolutely continuous with respect to $f_0^{(n)}$ in the sense that $f_0^{(n)}(\mathbf{x}) = 0 \Rightarrow f_1^{(n)}(\mathbf{x}) = 0$, then the optimal decision rule for the Bayesian or

Manuscript approved September 16, 1988.

Neyman-Pearson criterion [8] is given by the likelihood ratio test (LRT):

$$d(\mathbf{x}) = 1 \quad \text{iff} \quad \frac{f_1^{(n)}(\mathbf{x})}{f_0^{(n)}(\mathbf{x})} \geq \eta$$

where $\mathbf{x}$ is the observed vector, and the choice of the threshold η depends more specifically on the particular criterion and the details of the problem. If the observed random process is independent and identically distributed (iid) under either hypothesis, then only the marginal densities are involved, and the LRT becomes

$$d(\mathbf{x}) = 1 \quad \text{iff} \quad \prod_{i=1}^{n} \frac{f_1(x_i)}{f_0(x_i)} \geq \eta$$

where f_0 is the marginal density under H_0 and f_1 is the marginal density under H_1. Taking logarithms on each side, we may also write this in the form

$$d(\mathbf{x}) = 1 \quad \text{iff} \quad \sum_{i=1}^{n} \log \frac{f_1(x_i)}{f_0(x_i)} \geq \log \eta. \tag{1.2}$$

Because of the simple form of the LRT given by (1.2), this result has proven to be extremely useful in a practical sense whenever the processes can be assumed to be iid. However, if at least one of the processes is assumed to be dependent, the LRT might be of little practical value. Consider two such situations. First, it may be the case that one of the n-dimensional densities involved lacks a closed form expression, so that the LRT also lacks a closed form expression. Such is the generally the case for the Rayleigh distribution in Appendix A. In this situation the LRT is not implementable. Second, one may wish to implement a test which does not require an assumption on the particular form of the n-dimensional densities, such as is required for the LRT. For example, if one wants to design a test based on experimental data, then it is desirable to base such a test on the empirical marginal densities and possibly some of the lower order moments, since

the amount of data necessary to establish the n-dimensional empirical densities might be rather unmanageable. Since the LRT requires explicit knowledge of the n-dimensional densities, it is clear that the LRT is inappropriate in this case.

In the event that one chooses not to use an LRT, he may proceed by specifying a test structure and then attempt to determine the optimal test out the class of all tests which have that structure. A test structure which has appeared often in the literature is the following:

$$d(\mathbf{x}) = 1 \quad \text{iff} \quad T_n(\mathbf{x}) \in A \tag{1.3}$$

where

$$T_n(\mathbf{x}) = \sum_{i=1}^{n} \psi(x_i). \tag{1.4}$$

Here ψ is a Borel measurable function which will often be referred to as a nonlinearity, and A is a Borel subset of the real line which will be called the critical region. The test statistic T_n has been referred to in the literature as a zero-memory nonlinearity (ZNL). Note that in the case where the process is iid under either hypothesis, the log-LRT (1.2) has this form with $\psi(x) = \log[f_1(x)/f_0(x)]$ and $A = [\log \eta, \infty)$. Conversely, if the process is not iid under at least one of the hypotheses, then the LRT will involve memory, and consequently a memoryless decision rule will be *suboptimal.* Nevertheless, there are some advantages to using a memoryless decision rule, particularly when simplicity of implementation is important. In this thesis we consider in detail the use of various memoryless decision rules of the form (1.3) as applied to the discrimination problem (1.1) with the assmuption that the process is stationary under either hypothesis.

In order to implement a test of the form (1.3), one must specify the nonlinearity ψ and the critical region A. Most often, A will be an interval, such as the interval

$[\gamma, \infty)$ which involves a single threshold γ. In this case, one would probably proceed by specifying ψ first and then choosing the value of the threshold γ through simulation or actual testing to adjust the error probabilities to their desired values. Before determining the nonlinearity ψ, however, one must decide on a performance criterion. Then ψ will be chosen to be optimal with respect to this criterion. Although the most natural and useful criterion is that of the error probabilities, for a test of the form (1.3) one cannot in most cases obtain a closed form expression for the error probabilities, and thus another performance measure may be more useful. Such is the case for the results of this thesis, where we consider performance measures which involve the mean and asymptotic variance of the test statistic T_n under the two hypotheses. For stationary processes, the mean of T_n under H_i is given by

$$\mathrm{E}_i\, T_n(\mathbf{X}) = \mathrm{E}_i \sum_{j=1}^{n} \psi(X_j) = n\mathrm{E}_i\, \psi(X_1) \tag{1.5}$$

and the variance under H_i is

$$\begin{aligned}
\mathrm{Var}_i\, T_n(\mathbf{X}) &= \mathrm{E}_i\, T_n(\mathbf{X})^2 - [\mathrm{E}_i\, T(\mathbf{X})]^2 \\
&= \mathrm{E}_i \sum_{j=1}^{n}\sum_{k=1}^{n} \psi(X_j)\psi(X_k) - n^2\,[\mathrm{E}_i\, \psi(X_1)]^2 \\
&= \sum_{j=1}^{n} \mathrm{Var}_i\, \psi(X_j) + 2\sum_{j=1}^{n-1}\sum_{k=j+1}^{n} \mathrm{Cov}_i\,[\psi(X_j), \psi(X_k)] \\
&= n\mathrm{Var}_i\, \psi(X_1) + 2\sum_{j=1}^{n-1}\sum_{k=j+1}^{n} \mathrm{Cov}_i\,[\psi(X_1), \psi(X_{k-j+1})]\,.
\end{aligned} \tag{1.6}$$

In our notation, E_i, Var_i, and Cov_i denote, respectively, the expecation, variance, and covariance operations under hypothesis H_i. We now define two functionals $\mu_i(\psi)$ and $\sigma_i^2(\psi)$ which will appear in the work which follows. Define

$$\mu_i(\psi) = \mathrm{E}_i\, \psi(X_1) \tag{1.7}$$

and in cases where the sum exists, define

$$\sigma_i^2(\psi) = \mathrm{Var}_i\, \psi(X_i) + 2\sum_{j=1}^{\infty} \mathrm{Cov}_i\left[\psi(X_1), \psi(X_{j+1})\right]. \tag{1.8}$$

Thus we have $\mathrm{E}_i T_n = n\mu_i$, and we note also that if the sum in (1.8) converges, then $\frac{1}{n}\left|\mathrm{Var}_i\, T_n(\mathbf{X}) - n\sigma_i^2(\psi)\right| \to 0$ as $n \to \infty$ so that $\mathrm{Var}_i\, T_n(\mathbf{X}) \approx n\sigma_i^2(\psi)$ for large values of n. Thus $n\sigma_i^2$ is the asymptotic variance of T_n under hypothesis H_i. These functionals, or "moments," of the nonlinearity ψ have rather nice expressions in terms of the marginal and joint densities of the processes, and by considering performance measures involving these moments as opposed to the error probabilities, we shall find that the analysis becomes much more manageable. Note also that for such performance measures, the n-dimensional densities are not involved for $n > 2$.

In the chapters that follow, the performance measures which are derived are based on central limit theory; therefore it will be necessary to restrict the class of processes which will be considered. In particular, we desire that the processes involved demonstrate some kind of asymptotic independence so that central limit theory may be applied. The type of asymptotic independence which is appropriate for the work here is that which is defined by various *mixing* conditions. Let $\mathcal{F}_a^b$ denote the σ-field of events generated by $\{X_i, a \le i \le b\}$. Then the process $\{X_i\}$ is said to be *strong mixing* if there exists a sequence $\{\alpha_n\}$ such that $\alpha_n \to 0$ and

$$|P(A \cap B) - P(A)P(B)| \le \alpha_n \tag{1.9}$$

for any events $A \in \mathcal{F}_{-\infty}^k$, $B \in \mathcal{F}_{k+n}^{\infty}$. If it is also true that

$$|P(A \cap B) - P(A)P(B)| \le \phi_n P(B) \tag{1.10}$$

for some sequence $\{\phi_n\}$ with $\phi_n \to 0$, then the process is called *ϕ-mixing*. The ϕ-mixing condition clearly implies the strong mixing condition. Finally, define the process $\{X_i\}$ to be *m-dependent* if for every integer k we have that $\mathcal{F}_{-\infty}^{k}$ and $\mathcal{F}_{k+m+1}^{\infty}$ are independent. Note that m-dependence is a special case of ϕ-mixing with $\phi_n = 0$ for $n > m$. We include m-dependence because it is easier to work with analytically and because in certain situations it can approximate the ϕ-mixing condition well if m is sufficiently large. Because mixing conditions are defined in terms of the underlying σ-fields of events, the conditions are preserved by memoryless transformations, so that $\{g(X_i)\}$ will satisfy the same mixing condition as $\{X_i\}$, provided g is measurable. Central limit theorems have been proved for strong mixing and ϕ-mixing processes, and one such theorem is given here as Theorem 1.

Theorem 1. *Let $\{X_i\}$ be a stationary ϕ-mixing process with $\sum_{n=1}^{\infty} \phi_n^{1/2} < \infty$ and let ψ be a measurable real-valued function such that $\mathrm{E}|\psi(X_1)| < \infty$ and $\mathrm{E}\psi(X_1)^2 < \infty$. Then the series in (1.8) converges absolutely and $[T_n(\mathbf{X}) - n\mu(\psi)]/\sqrt{n\sigma^2(\psi)}$ converges in distribution to a standard normal random variable (having zero mean and unit variance), provided $\sigma^2(\psi) > 0$.*

For a proof, see [2] and [4]. In Chapters 2 and 3, we shall assume the m-dependent or ϕ-mixing condition and make reference to Theorem 1. In Chapter 4 we shall assume the strong mixing condition and shall also state there a central limit theorem for strong mixing processes.

A performance measure which has received a lot of attention in the literature is that of the efficacy of a test, which is based on the concept of asymptotic relative

efficiency (ARE). In order to define the efficacy of a test, consider the following problem:

$$\begin{aligned} H_0&: \quad X_i = N_i \\ H_1&: \quad X_i = N_i + \theta \end{aligned} \tag{1.11}$$

where the process $\{N_i\}$ is a stationary process which represents a noise process. Thus under H_0 we observe strictly noise while under H_1 we observe a constant signal θ plus noise. Let $f_N^{(n)}$ denote the n dimensional density of the noise process. Note that this situation is a special case of the problem (1.1) with $f_0^{(n)}(\mathbf{x}) = f_N^{(n)}(\mathbf{x})$ and $f_1^{(n)}(\mathbf{x}) = f_N^{(n)}(\mathbf{x}-\theta)$. If we consider a test involving the test statistic $T_n^{(1)}$, then for a given signal strength θ and fixed error probabilities α, β under H_0 and H_1, respectively, a minimum sample size n_1 is required. Under similar conditions but with a different test statistic $T_n^{(2)}$ a sample size n_2 is required. The ARE of $T^{(2)}$ with respect to $T^{(1)}$ may now be defined as the limit of the ratio n_1/n_2 as $\theta \to 0$ and α and β remain fixed. The efficacy of $T^{(1)}$ is defined to be

$$\eta_1 = \lim_{n\to\infty} \frac{\left[\frac{d}{d\theta}\mathrm{E}_\theta T^{(1)}\right]^2_{\theta=0}}{n\mathrm{Var}_0 T^{(1)}} \tag{1.12}$$

if the limit exists. Here the subscripts for the expectation and variance operators denote the value of θ under which the operations are performed; e.g. Var_0 denotes the variance under H_0. The importance of the efficacy stems from the Pitman-Noether theorem, which states that if certain conditions are satisfied individually by $T^{(1)}$ and $T^{(2)}$, then the ARE of $T^{(2)}$ with respect to $T^{(1)}$ is equal to the ratio η_2/η_1 of the two efficacies. Hence $T^{(1)}$ may be considered a better test than $T^{(2)}$ under the ARE criterion if $\eta_1 > \eta_2$. The conditions on $T^{(1)}$ which are necessary for the Pitman-Noether theorem are as follows:

(i) $\dfrac{d}{d\theta}\mathrm{E}_\theta T^{(1)}\big|_{\theta=0} > 0$

(ii) $\eta_1 > 0$

(iii) $\lim_{n\to\infty} \dfrac{\left[\frac{d}{d\theta}\mathrm{E}_\theta T^{(1)}\right]^2_{\theta=\theta_n}}{\left[\frac{d}{d\theta}\mathrm{E}_\theta T^{(1)}\right]^2_{\theta=0}} = \lim_{n\to\infty} \dfrac{\mathrm{Var}_{\theta_n} T^{(1)}}{\mathrm{Var}_0 T^{(1)}} = 1$ where $\theta_n = K/\sqrt{n}$ for some constant K.

(iv) $\left[T_n^{(1)} - \mathrm{E}_\theta T_n^{(1)}\right] / \sqrt{\mathrm{Var}_\theta T_n^{(1)}}$ converges in distribution to a standard normal random variable as $n \to \infty$ for all $\theta \in [0, \tilde{\theta})$.

Similar conditions are required of $T^{(2)}$. The condition (iv) requires that the test statistic $T^{(1)}$ be asymptotically normal. In cases where the noise process is iid, it is straightforward to apply a central limit theorem to show that the condition (iv) holds for the test statistic in (1.4), and Miller and Thomas [11] have derived the nonlinearity ψ which maximizes the efficacy in such a situation. They also generalize to the case of a nonconstant signal involving the test statistic $T_n(\mathbf{x}) = \sum_{i=1}^n \psi_i(x_i)$ for which the nonlinearity varies with time. In Poor and Thomas [1], the optimal nonlinearity is derived, still with respect to the efficacy performance measure, for the case where the noise process is m-dependent. Halverson and Wise [2] show how to correctly extend this result to the more general case of ϕ-mixing noise. In either of these latter two cases, Theorem 1 is required in order to demonstrate that the test statistic is asymptotically normal.

In this thesis we shall not confine ourselves to the weak signal in noise problem, but we shall consider the more general problem of discrimination (1.1), with the assumption that the observed process is stationary and satisfies a mixing condition under either hypothesis. For this type of problem the efficacy performance measure is no longer appropriate. We shall therefore derive the appropriate performance measures. These performance measures, which are derived under different problem formulations, are all of the form

$$\frac{(\mu_1 - \mu_0)^2}{\|(\sigma_0, \sigma_1)\|^2}, \tag{1.13}$$

where the denominator is the square of some "norm" of the vector (σ_0, σ_1), and where μ_i and σ_i are defined by (1.7) and (1.8). Thus these performance measures are functionals of the nonlinearities. In Chapter 2 we consider the problem (1.1) under a Neyman-Pearson formulation. Thus if P_i denotes the probability of error when H_i is true, then the formulation considered here is to minimize P_1 subject to the constraint that $P_0 \leq \alpha$. For this situation, it will be shown that the optimal receiver for large sample sizes (that is, in an asymptotic sense) is such that it maximizes the performance measure

$$S_1 = \frac{(\mu_1 - \mu_0)^2}{\sigma_1^2}. \tag{1.14}$$

For the reverse situation where P_0 is minimized subject to $P_1 \leq \alpha$, the performance measure is

$$S_0 = \frac{(\mu_1 - \mu_0)^2}{\sigma_0^2}. \tag{1.15}$$

In each of these two performance measures, the "norm" in the denominator is $\|(\sigma_0, \sigma_1)\| = \sigma_i$, which, technically speaking, is a pseudo-norm, since $\|(\sigma_0, \sigma_1)\| = 0$ does not necessarily imply that $(\sigma_0, \sigma_1) = (0, 0)$. Observe the similarity of the performance measure S_1 to the efficacy measure (1.12). This similarity arises from the fact that both performance measures are aymptotic performance measures based on central limit theory. For the efficacy, however, the assumptions (i)–(iv) are necessary, whereas fewer assumptions are necessary to justify the use of S_1. Also in Chapter 2, the nonlinearity which maximizes the performance measure $^{c}_{.}$ is shown to satisfy a Fredholm integral equation of the second kind. It is also shown how the integral equation can be solved using Hilbert-Schmidt theory. In Chapter 3, we consider again the problem (1.1), this time under a minimax formulation; that is, we desire to minimize the maximum of P_0 and P_1. It will be shown

that the optimum test statistic is one which maximizes the performance measure

$$S_2 = \frac{(\mu_1 - \mu_0)^2}{(\sigma_0 + \sigma_1)^2}. \tag{1.16}$$

The nonlinearity which is optimal for this performance measure is shown to satisfy a nonlinear integral equation for which a closed form solution cannot be given; however, it is shown how the solution can be obtained numerically using an iterative procedure. By modifying the minimax formulation slightly, it is shown that one can also derive the performance measure

$$S_3 = \frac{(\mu_1 - \mu_0)^2}{\sigma_0^2 + \sigma_1^2}, \tag{1.17}$$

which has the Euclidean norm in the denominator. It will be shown that the maximization of S_3 leads to a linear integral equation. In Chapter 4, the issue of robustness is addressed. The approach is that of game theory, or minimax theory, where one tries to design the optimal receiver to match the worst case densities chosen out of uncertainty classes. Results are given here for the performance measures S_1 (and consequently S_0 as well) and S_3. In Chapter 5, the theory is applied to the problem of discrimination between a Rayleigh density and a lognormal density, where strong correlation is present. The nonlinearity which maximizes each of the performance measures is computed numerically, and the performance results from computer simulations are presented. The simulation results are compared to the results for the receiver which is designed under the assumption that the processes are iid. Chapter 5 also contains a discussion of the results of the thesis.

CHAPTER 2

THE NEYMAN-PEARSON FORMULATION

2.1 The performance measure S_1

In this chapter, we consider in detail the hypothesis testing problem (1.1) under a Neyman-Pearson formulation. Our informal statement of the problem is the following:

$$\text{minimize} \quad P_1 \tag{2.1}$$

$$\text{subject to} \quad P_0 \leq \alpha$$

where P_i denotes the probability of error when H_i is true. The reason that the statement of the problem (2.1) is informal is because it depends implicitly on the sample size n, and although we are interested in tests with a fixed sample size, we do not wish to specify n before we consider the problem (2.1). When we speak of a test or decision rule, we shall actually mean a family of decision rules—one for each n—and in comparing different tests, we shall not explicitly mention a particular value of n. Since for any reasonable test $P_1 \to 0$ as $n \to \infty$, we may state our problem more accurately in this way: considering all level α tests (i.e. $P_0 \leq \alpha$ for all n), find the test for which the rate of convergence of P_1 to 0 is fastest. We can see now that if for some test $d^{(1)}$ the rate of

convergence is faster than the rate for another test $d^{(2)}$, then there is an integer N such that $d^{(1)}$ is better than $d^{(2)}$ in the sense implied by (2.1) whenever the sample size n is greater than N. In this section we shall derive a performance measure S_1 which specifies (approximately) the rate at which P_1 converges to 0, and the connection between this performance measure and the Neyman-Pearson problem (2.1) should be clear.

We will restrict our attention to only those decision rules of the form (1.3) with the assumption that the test statistic T_n is asymptotically normal under either hypothesis. Thus under hypothesis H_i we assume that there exist constants μ_i and $\sigma_i^2 > 0$ such that $(T_n - n\mu_i)/\sqrt{n\sigma_i^2}$ converges in distribution to a standard normal random variable as $n \to \infty$. In the case that the test statistic has the form (1.4) and the conditions of Theorem 1 are satisfied, the constants μ_i and σ_i^2 are given by (1.7) and (1.8), respectively. With this assumption, then, for large values of n the distribution of the test statistic is approximately normal with mean $n\mu_i$ and variance $n\sigma_i^2$ when H_i is true. Taking a heuristic approach, one can use this knowledge to choose the critical region A by considering the decision rule (1.3) to be equivalent to an LRT between two Gaussian densities. In our case, the two Gaussian densities are

$$\begin{aligned} \varphi_0(t) &= \frac{1}{\sqrt{2\pi n}\sigma_0} \exp\left\{-\frac{(t-n\mu_0)^2}{2n\sigma_0^2}\right\} \\ \varphi_1(t) &= \frac{1}{\sqrt{2\pi n}\sigma_1} \exp\left\{-\frac{(t-n\mu_1)^2}{2n\sigma_1^2}\right\} \end{aligned} \tag{2.2}$$

and the log-LRT is given by

$$d(\mathbf{x}) = 1 \qquad \text{iff} \qquad \log\left[\frac{\varphi_1(T_n)}{\varphi_0(T_n)}\right] \geq n\eta,$$

or equivalently,

$d(\mathbf{x}) = 1$ iff

$$\frac{1}{n}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right) T_n^2 - 2\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right) T_n + n\left(\frac{\mu_0^2}{\sigma_0^2} - \frac{\mu_1^2}{\sigma_1^2} - \gamma\right) \geq 0 \tag{2.3}$$

where $\gamma = 2\eta - (2/n)\log(\sigma_0/\sigma_1)$. We can assume without loss of generality that $\mu_0 < \mu_1$: the case of $\mu_1 = \mu_0$ is unlikely to occur in practice and will not be considered, and the case of $\mu_1 < \mu_0$ follows the same procedure with the appropriate sign change. Assume first that $\sigma_0^2 = \sigma_1^2$. If this is true, the expression on the left side of (2.3) is a linear function of T_n, and it is easy to see that the log-LRT has the form (1.3) with $A = [n\gamma', \infty)$, where $n\gamma'$ is the root of the linear function in (2.3). The error probabilities are then given approximately by

$$\begin{aligned} P_0 &= \Phi\left[-\sqrt{n}\frac{\gamma' - \mu_0}{\sigma_0}\right] \\ P_1 &= \Phi\left[\sqrt{n}\frac{\gamma' - \mu_1}{\sigma_1}\right] \end{aligned} \tag{2.4}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Now in order to have $P_0 = \alpha$, we must take $\gamma' = (\sigma_0/\sqrt{n})\Phi^{-1}(\alpha) + \mu_0$, and substituting for γ' in the expression for P_1 we obtain

$$P_1 = \Phi\left[\Phi^{-1}(\alpha) - \sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_1}\right]. \tag{2.5}$$

Now $\Phi(x)$ is an increasing function of x and so to minimize P_1, it is necessary to make the argument of Φ in (2.5) as small as possible; that is, to make the argument large in magnitude and negative. The term $-\sqrt{n}(\mu_1 - \mu_0)/\sigma_1$ is negative, since we are assuming that $\mu_0 < \mu_1$, and it increases in magnitude as n increases, so that $P_1 \to 0$. The quantity $(\mu_1 - \mu_0)/\sigma_1$ determines the rate at which P_1 goes to zero, and we can see that the best

asymptotic performance results when this quantity is maximized. Since it is positive, it is also clear that maximizing $(\mu_1 - \mu_0)/\sigma_1$ is equivalent to maximizing the quantity

$$S_1 = \frac{(\mu_1 - \mu_0)^2}{\sigma_1^2}. \tag{2.6}$$

Assuming now that $\sigma_0^2 > \sigma_1^2$, we will obtain the same performance measure; the case of $\sigma_0^2 < \sigma_1^2$ will not be considered since the analysis parallels the case of $\sigma_0^2 > \sigma_1^2$ and the same result is obtained. We see that the test (2.3) is identical to the test (1.3) if $A = [n\gamma_1, n\gamma_2]$, where $n\gamma_1$ and $n\gamma_2$ are the roots of the quadratic in (2.3). The quantities γ_1, γ_2 are given by the expressions

$$\begin{aligned} \gamma_1 &= \frac{\mu_1\sigma_0^2 - \mu_0\sigma_1^2 - \sigma_0\sigma_1\sqrt{(\mu_1 - \mu_0)^2 - \gamma(\sigma_0^2 - \sigma_1^2)}}{\sigma_0^2 - \sigma_1^2} \\ \gamma_2 &= \frac{\mu_1\sigma_0^2 - \mu_0\sigma_1^2 + \sigma_0\sigma_1\sqrt{(\mu_1 - \mu_0)^2 - \gamma(\sigma_0^2 - \sigma_1^2)}}{\sigma_0^2 - \sigma_1^2} \end{aligned} \tag{2.7}$$

and error probabilities are given approximately by

$$\begin{aligned} P_0 &= \Phi\left[\sqrt{n}\frac{\gamma_2 - \mu_0}{\sigma_0}\right] - \Phi\left[\sqrt{n}\frac{\gamma_1 - \mu_0}{\sigma_0}\right] \\ P_1 &= \Phi\left[-\sqrt{n}\frac{\gamma_2 - \mu_1}{\sigma_1}\right] + \Phi\left[\sqrt{n}\frac{\gamma_1 - \mu_1}{\sigma_1}\right]. \end{aligned} \tag{2.8}$$

Substituting for γ_1 and γ_2 yields

$$\begin{aligned} P_0 &= \Phi\left[\sqrt{n}\frac{\sigma_0(\mu_1 - \mu_0) + \sigma_1 v(\gamma)}{\sigma_0^2 - \sigma_1^2}\right] - \Phi\left[\sqrt{n}\frac{\sigma_0(\mu_1 - \mu_0) - \sigma_1 v(\gamma)}{\sigma_0^2 - \sigma_1^2}\right] \\ P_1 &= \Phi\left[-\sqrt{n}\frac{\sigma_1(\mu_1 - \mu_0) + \sigma_0 v(\gamma)}{\sigma_0^2 - \sigma_1^2}\right] + \Phi\left[\sqrt{n}\frac{\sigma_1(\mu_1 - \mu_0) - \sigma_0 v(\gamma)}{\sigma_0^2 - \sigma_1^2}\right] \end{aligned} \tag{2.9}$$

where $v(\gamma) = \sqrt{(\mu_1 - \mu_0)^2 - \gamma(\sigma_0^2 - \sigma_1^2)}$. Thus we have the approximate error probabilities given as functions of a parameter γ.

For situations where the error probabilities are relatively small, little is to be gained by preferring a two threshold test to a single threshold test. This is the gist of

Proposition 2 below. In order to make the proposition precise, it is necessary to assume that T_n is truly Gaussian and not merely an approximation.

Proposition 2. *Let T_n be a test statistic which has the distribution $\mathcal{N}(n\mu_i, n\sigma_i^2)$ under hypothesis H_i for $i = 0, 1$, with $\sigma_0^2 > \sigma_1^2$. Let $P_0^{(2)}$ and $P_1^{(2)}$ denote the error probabilities for the decision rule $d^{(2)}$ of the form (1.3) with $A = [n\gamma_1, n\gamma_2]$ where γ_1 and γ_2 are given by (2.7). Let $P_0^{(1)}$ and $P_1^{(1)}$ denote the error probabilities for the decision rule $d^{(1)}$, also of the form (1.3), but with $A = [n\gamma_1, \infty)$. Assume that the thresholds are chosen for each sample size n so that $P_0^{(2)} = \alpha$. Then as $n \to \infty$, we have*

$$\left| \frac{P_1^{(2)} - P_1^{(1)}}{P_1^{(2)}} \right| \to 0.$$

Proof. Define

$$\nu_1 = \frac{\sigma_1(\mu_1 - \mu_0) + \sigma_0 v(\gamma)}{\sigma_0^2 - \sigma_1^2} \qquad \lambda_1 = \frac{\sigma_0(\mu_1 - \mu_0) + \sigma_1 v(\gamma)}{\sigma_0^2 - \sigma_1^2}$$

$$\nu_2 = -\frac{\sigma_1(\mu_1 - \mu_0) - \sigma_0 v(\gamma)}{\sigma_0^2 - \sigma_1^2} \qquad \lambda_2 = \frac{\sigma_0(\mu_1 - \mu_0) - \sigma_1 v(\gamma)}{\sigma_0^2 - \sigma_1^2}.$$

Then from (2.9) we have

$$P_1^{(2)} = \Phi(-\sqrt{n}\nu_1) + \Phi(-\sqrt{n}\nu_2)$$

$$P_1^{(1)} = \Phi(-\sqrt{n}\nu_2).$$

and

$$P_0^{(2)} = \Phi(\sqrt{n}\lambda_1) - \Phi(\sqrt{n}\lambda_2). \tag{2.10}$$

Since $\lambda_1 > \sigma_0(\mu_1 - \mu_0)/(\sigma_0^2 - \sigma_1^2) > 0$, the first term of (2.10) converges to 1 as $n \to \infty$. But $P_0^{(2)} = \alpha$, so the second term converges to $1 - \alpha$. Therefore $\lambda_2 \to 0$. This implies that $v(\gamma) \to (\sigma_0/\sigma_1)(\mu_1 - \mu_0)$ and thus

$$\nu_1 \to \frac{(\sigma_0^2 + \sigma_1^2)(\mu_1 - \mu_0)}{\sigma_1(\sigma_0^2 - \sigma_1^2)} > 0, \qquad \nu_2 \to \frac{\mu_1 - \mu_0}{\sigma_1} > 0.$$

The proposition will follow if we can show that

$$\frac{\Phi(-\sqrt{n}\nu_1)}{\Phi(-\sqrt{n}\nu_2)} \to 0.$$

To do this, use the inequalities [8,p.39]

$$-\frac{1}{\sqrt{2\pi}x}\left(1-\frac{1}{x^2}\right)e^{-x^2/2} < \Phi(x) < -\frac{1}{\sqrt{2\pi}x}e^{-x^2/2}$$

which are valid for $x < 0$. Thus we have

$$\frac{\Phi(-\sqrt{n}\nu_1)}{\Phi(-\sqrt{n}\nu_2)} < \frac{\nu_1}{\nu_2}\left(\frac{n\nu_2^2}{n\nu_2^2-1}\right)\exp\left[-\frac{n}{2}(\nu_1^2-\nu_2^2)\right] \to 0.$$

□

Because the test statistics which we consider are only approximately normal, this proposition does not directly apply, and it has been presented to provide a heuristic argument for a single threshold test. In fact, the proof of the proposition depends in a crucial way on the tail behavior of the distributions, and for a series which converges under a central limit theorem, convergence is usually slowest in the tail region. In most practical situations, however, the realizations of the test statistic under H_i tend to pile up around $n\mu_i$, so that if $n(\mu_1 - \mu_0)$ is large then a two threshold test offers no advantage over a single threshold test. Thus we may justify a single threshold test not only on a heuristic basis but also from practical considerations.

We are now ready to derive the performance measure S_1 for the case where $\sigma_0^2 > \sigma_1^2$ and where a single threshold test of the form (1.3) is used. With $A = [n\gamma, \infty)$, we have the error probability under H_0 given approximately by

$$P_0 = \Phi\left[-\sqrt{n}\frac{\gamma-\mu_0}{\sigma_0}\right].$$

If we set $P_0 = \alpha$ then we find the value of γ to be

$$\gamma = -\frac{\sigma_0}{\sqrt{n}}\Phi^{-1}(\alpha) + \mu_0.$$

On the other hand, the error probability under H_1 is given approximately by

$$P_1 = \Phi\left[\sqrt{n}\frac{\gamma - \mu_1}{\sigma_1}\right] = \Phi\left[-\frac{\sigma_0}{\sigma_1}\Phi^{-1}(\alpha) - \sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_1}\right] \tag{2.11}$$

where the last equality in (2.11) is obtained by making the substitution for γ. Now the quantity σ_0/σ_1 does not depend on the sample size n. Therefore the quantity $(\mu_1 - \mu_0)/\sigma_1$ again determines the rate at which P_1 converges to zero. By the same reasoning as before, then, we see that the best asymptotic performance results when S_1, as given by (2.6), is maximized.

The results obtained in this section do not depend on the form of the test statistic T_n, only on the assumption that there exist the constants μ_0, μ_1, σ_0, and σ_1 such that $(T_n - n\mu_i)/\sqrt{n\sigma_i^2}$ converges in distribution to a standard normal random variable when H_i is true. In the remainder of this chapter, we restrict our attention to test statistics of the form $T_n = \sum_{i=1}^{n} g(x_i)$ where the conditions of Theorem 1 hold. Thus the "moments" μ_0, μ_1, σ_0^2, and σ_1^2 are given by (1.7) and (1.8) and the performance measure S_1 becomes a functional of g.

2.2 The Optimal Nonlinearity

In the first section of this chapter we showed heuristically that the best test statistic in the asymptotic sense for the Neyman-Pearson problem (2.1) is that for which the performance measure S_1 is maximized. In this section, we will consider the following optimization problem

$$\text{maximize} \qquad S_1(g) = \frac{\left[\mu_1(g) - \mu_0(g)\right]^2}{\sigma_1^2(g)} \tag{2.12}$$

subject to the constraints that $\mathrm{E}_i\, g^2(X_1) < \infty$ for $i = 0, 1$. Thus if g_1 solves (2.12), then the test statistic $T_n = \sum g_1(x_i)$ is optimal in the sense of Section 2.1 over the class of all memoryless test statistics (1.4). In the next section conditions are given which guarantee that the nonlinearity g_1 derived in this section satisfies the constraint $\mathrm{E}_1\, g_1^2(X_1) < \infty$. In order to have $\mathrm{E}_0\, g_1^2(X_1) < \infty$, then, it is sufficient to require that $f_0(x)/f_1(x)$ be bounded for all x, and we shall make this assumption. We shall also take this condition to mean that $f_1(x) = 0 \;\Rightarrow\; f_0(x) = 0$ as well. We assume that g and all the densities involved are continuous so that we can apply the classical techniques from the calculus of variations. Naturally we will have to assume that the conditions of Theorem 1 are satisfied under each hypothesis, so that the test statistic T_n satisfies our assumption of asymptotic normality. In this section, however, we shall require the more stringent condition that the observed process be m-dependent under either hypothesis. Observe that for an m-dependent process the expression for $\sigma_i^2(g)$ as given by (1.8) becomes

$$\sigma_i^2(g) = \mathrm{E}_i\, g^2(X_1) + \sum_{j=1}^{m} \mathrm{E}_i\, g(X_1) g(X_{j+1}) - (2m+1)\,[\mathrm{E}_i\, g(X_1)]^2 \,. \tag{2.13}$$

At the end of this chapter, we shall discuss the extension of the m-dependent results to the more general case of ϕ-mixing processes, and show that for all practical purposes, ϕ-mixing processes can be approximated by m-dependent ones.

In the remainder of this section, we derive an integral equation for which the optimal nonlinearity g_1 is a solution. We begin by observing that the value of S_1 is unchanged when g is multiplied by a constant, hence we can maximize $(\mu_1 - \mu_0)^2$ with σ_1^2 held constant. But under our assumption that $\mu_1 - \mu_0 > 0$, maximizing $(\mu_1 - \mu_0)^2$ is equivalent to maximizing $\mu_1 - \mu_0$. We will therefore introduce a Lagrange multiplier λ

and consider the problem

$$\text{maximize} \quad \mu_1 - \mu_0 - \lambda\sigma_1^2. \tag{2.14}$$

Now define $J(g) = \mu_1 - \mu_0 - \lambda\sigma_1^2$. A necessary condition for g to maximize $J(\cdot)$ is that the Gâteaux variation of J evaluated at g vanish. Thus we must have $\frac{d}{d\epsilon} J(g + \epsilon\delta g)\big|_{\epsilon=0} = 0$, where δg is an arbitrary continuous function satisfying $\mathrm{E}_i\, \delta g^2(X_1) < \infty$ for $i = 0, 1$. Since

$$\begin{aligned} J(g + \epsilon\delta g) = J(g) + \epsilon\Big[&\mathrm{E}_1\, \delta g(X_1) - \mathrm{E}_0\, \delta g(X_1) - 2\lambda\Big\{\mathrm{E}_1\, g(X_1)\delta g(X_1) \\ &+ \sum_{j=1}^{m}\left[\mathrm{E}_1\, g(X_1)\delta g(X_{j+1}) + \mathrm{E}_1\, g(X_{j+1})\delta g(X_1)\right] \\ &- (2m+1)\mathrm{E}_1\, g(X_1)\mathrm{E}_1\, \delta g(X_1)\Big\}\Big] + \epsilon^2\left[-\lambda\sigma_1^2(\delta g)\right], \end{aligned} \tag{2.15}$$

which is a quadratic function in ϵ, the Gâteaux variation is given by the coefficient of ϵ. Denote by f_i and f_i^j the densities of X_1 and (X_1, X_{j+1}), respectively, under H_i. If we introduce these densities into the above expression and set the coefficient of ϵ equal to 0, we obtain

$$\begin{aligned} 0 = \int \delta g(x)\Big\{&f_1(x) - f_0(x) - 2\lambda\Big[f_1(x)g(x) \\ &+ \int\Big(\sum_{j=1}^{m}\left[f_1^j(x,y) + f_1^j(y,x)\right] - (2m+1)f_1(x)f_1(y)\Big)g(y)dy\Big]\Big\}dx. \end{aligned} \tag{2.16}$$

The right-hand side of (2.16) will be zero for arbitrary δg iff the quantity in the braces is identically zero. Setting this expression equal to zero, we easily derive the following integral equation:

$$2\lambda g(x) = \frac{f_1(x) - f_0(x)}{f_1(x)} + 2\lambda \int K_1(x,y)g(y)dy \tag{2.17}$$

with the kernel K_1 given by

$$K_1(x,y) = (2m+1)f_1(y) - \frac{1}{f_1(x)}\sum_{j=1}^{m}\left[f_1^j(x,y) + f_1^j(y,x)\right]. \tag{2.18}$$

In the next section we will discuss the conditions which guarantee the existence of a solution to this integral equation. For now, we note that in order to divide by $f_1(x)$ as we have done, we require the absolute continuity condition $f_1(x) = 0 \Rightarrow f_0(x) = 0$.

We have derived the integral equation (2.17) as a necessary condition for g to solve the maximization problem (2.14). However, if g solves (2.17) then $J(g+\epsilon\delta g) = J(g) - \epsilon^2\lambda\sigma_1^2$ so that $J(g) \geq J(g+\epsilon\delta g)$ provided $\lambda > 0$. Thus if $\lambda > 0$, the condition that g solve the integral equation (2.17) is also sufficient for g to solve the maximization problem. Observing the form of (2.17) it is obvious that λ determines the scaling of g, thus we may take an arbitrary (positive) value for λ. In the analysis that follows, it will be convenient to take $\lambda = \frac{1}{2}$, and with this value the integral equation (2.17) becomes

$$g(x) = \frac{f_1(x) - f_0(x)}{f_1(x)} + \int K_1(x,y)g(y)dy. \tag{2.19}$$

If we make the substitution $g(x) = h(x)/\sqrt{f_1(x)}$, the integral equation (2.19) becomes

$$h(x) = \frac{f_1(x) - f_0(x)}{\sqrt{f_1(x)}} + \int K_1^*(x,y)h(y)dy \tag{2.20}$$

which has the symmetric kernel

$$K_1^*(x,y) = (2m+1)\left[f_1(x)f_1(y)\right]^{\frac{1}{2}} - \left[f_1(x)f_1(y)\right]^{-\frac{1}{2}}\sum_{j=1}^{m}\left[f_1^j(x,y) + f_1^j(y,x)\right]. \tag{2.21}$$

Our purpose for making the substitution for g to get the equation (2.20) is to permit us to apply the Hilbert-Schmidt theory for symmetric integral equations. This is the topic of the next section.

Consider now $\sigma_1^2(g_1)$, which we may write in the form

$$\sigma_1^2(g_1) = \int g_1^2(x) f_1(x) dx + \sum_{j=1}^{m} \iint g_1(x) g_1(y) [f_1^j(x,y) + f_1^j(y,x)] dx dy$$

$$- (2m+1) \iint g_1(x) g_1(y) f_1(x) f_1(y) dx dy$$

$$= \int g_1(x) f_1(x) \Big[g_1(x) + \int \Big\{ \frac{1}{f_1(x)} \sum_{j=1}^{m} [f_1^j(x,y) + f_1^j(y,x)]$$

$$- (2m+1) f_1(y) \Big\} g_1(y) dy \Big] dx$$

$$= \int g_1(x) f_1(x) [g_1(x) - \int K_1(x,y) g_1(y) dy] dx$$

From this expression it can be seen that if g_1 solves the integral equation (2.19), then $\sigma_1^2(g_1) = \mu_1(g_1) - \mu_0(g_1)$, and thus $S_1(g_1) = \mu_1(g_1) - \mu_0(g_1)$ is the optimal value of S_1. We summarize the results of this section in the following theorem.

Theorem 3. *If the process $\{X_i\}$ is m-dependent under both H_0 and H_1, then a sufficient condition for g_1 to maximize S_1 is that g_1 solve the integral equation (2.19). Furthermore, if g_1 solves (2.19) then $S_1(g_1) = \mu_1(g_1) - \mu_0(g_1)$.*

2.3 The Solution of the Integral Equation

To apply the theory of Fredholm equations we require the following two conditions:

$$(a) \quad \int \frac{[f_1(x) - f_0(x)]^2}{f_1(x)} dx < \infty$$

$$(b) \quad \iint |K_1^*(x,y)|^2 dx dy < \infty.$$

Under the assumption that $f_0(x)/f_1(x)$ is bounded, the condition (a) follows easily. To show condition (b), we assume that the densities $f_1^j(x,y), j = 1, ..., m$ have the diagonal

expansion [5]

$$f_1^j(x,y) = f_1(x)f_1(y)\sum_{n=1}^{\infty} a_n^{(j)}\theta_n(x)\theta_n(y). \tag{2.22}$$

where the functions $\{\theta_n\}$ are orthonormal in the sense that $\int \theta_m(x)\theta_n(x)f_1(x)dx = \delta_{mn}$. Some examples of densities which are known to have such an expansion are the Gaussian and gamma densities. Consider now the terms in the expansion of $|\tilde{K}^*|^2$. We examine only the terms of the form $f_1^j(x,y)f_1^k(x,y)/[f_1(x)f_1(y)]$, the other terms being more obviously integrable. If we introduce the expansion (2.22) and apply the orthogonality relation, we have

$$\iint \frac{f_1^j(x,y)f_1^k(x,y)}{f_1(x)f_1(y)}dxdy = \sum_{n=0}^{\infty} a_n^{(j)}a_n^{(k)} < \infty$$

from which the condition (b) follows. Conditions (a) and (b) are sufficient to guarantee that the solution $h_1(x) = \sqrt{f_1(x)}g_1(x)$, if it exists, is square integrable, and this in turn implies that $\mathrm{E}_1\, g_1^2(X_1) < \infty$.

In the iid case, the kernel K_1^* reduces to $[f_1(x)f_1(y)]^{\frac{1}{2}}$, and it is easy to verify the solution

$$h(x) = \frac{cf_1(x) - f_0(x)}{\sqrt{f_1(x)}}$$

where c is an arbitrary constant. Note that the absolute term of the integral equation is of this form when $c = 1$. We may therefore define h_{iid} by

$$h_{\mathrm{iid}}(x) = \frac{f_1(x) - f_0(x)}{\sqrt{f_1(x)}} \tag{2.23}$$

and write the integral equation as

$$h(x) = h_{\mathrm{iid}}(x) + \int K_1^*(x,y)h(y)dy. \tag{2.24}$$

When the process is not iid, we can still solve the integral equation using the Hilbert-Schmidt theory, provided we can find the eigenvalues and eigenvectors of the kernel. According to the theory, a unique solution exists provided the conditions (a) and (b) above hold, and provided +1 is not an eigenvalue of the kernel. If 1 is an eigenvalue, a (non-uniqne) solution still exists, provided the absolute term h_{iid} is orthogonal to every eigenvector corresponding to the eigenvalue 1. We shall see that 1 is an eigenvalue of the kernel K_1^*, but that in most cases a solution still exists.

If we assume that the densities f_1^j, $j = 1, \ldots, m$ have the expansion (2.22) and we introduce this expansion into the kernel, we have

$$K_1^*(x,y) = \sqrt{f_1(x)f_1(y)}\left[(2m+1) - 2\sum_{n=0}^{\infty}\left(\sum_{j=1}^{m} a_n^{(j)}\right)\theta_n(x)\theta_n(y)\right]. \tag{2.25}$$

Usually we will have $\theta_0(x) \equiv 1$ for such an expansion, and in such cases K_1^* will have eigenvalues $\{\lambda_n\}$ and eigenvectors $\{\phi_n\}$ given by

$$\lambda_0 = 1$$

$$\lambda_n = -2\sum_{j=1}^{m} a_n^{(j)} \qquad (n \geq 1)$$

$$\phi_n = \sqrt{f_1}\theta_n \qquad (n \geq 0).$$

Since $\lambda_0 = 1$, we must verify that h_{iid} is orthogonal to $\phi_0 = \sqrt{f_1}$, which is trivial:

$$\int h_{\text{iid}}(x)\phi_0(x)dx = \int f_1(x) - f_0(x)dx = 0.$$

If $\lambda_n \neq 1, n \geq 1$, we have the solution

$$h(x) = h_{\text{iid}}(x) + \sum_{n\geq 1} \frac{\lambda_n c_n}{1-\lambda_n}\phi_n(x) + c\phi_0(x)$$

$$= h_{\text{iid}}(x) + \sum_{n\geq 1} \frac{\lambda_n c_n}{1-\lambda_n}\theta_n(x)\sqrt{f_1(x)} + c\sqrt{f_1(x)}$$

with

$$c_n = \int h_{\text{iid}}(x)\phi_n(x)dx = \int [f_1(x) - f_0(x)]\,\theta_n(x)dx \tag{2.26}$$

and c an arbitrary constant. Therefore, the nonlinearity g (with $c = -1$) is given by

$$g(x) = -\frac{f_0(x)}{f_1(x)} + \sum_{n\geq 1} \frac{\lambda_n c_n}{1-\lambda_n}\theta_n(x). \tag{2.27}$$

2.4 Extension to ϕ-Mixing Processes

In this section we consider again the optimization problem (2.12) under the more general case where the processes are assumed to be ϕ-mixing. We will use a compactness argument to prove that the optimization problem has a solution g_1 and then show that if $g_1^{(m)}$ solves the integral equation (2.19) then the sequence $S_1(g_1^{(m)})$ converges to the optimal value $S_1(g_1)$ as $m \to \infty$. Obviously, similar results hold for the performance measure S_0.

First, let us define some new symbols. Let $L^2(f_1)$ denote the Hilbert space consisting of the Borel functions g such that $\int g^2(x)f_1(x)dx < \infty$ with the inner product $\langle g,h\rangle = \int g(x)h(x)f_1(x)dx$ and norm $\|g\| = \left[\int g^2(x)f_1(x)dx\right]^{\frac{1}{2}}$. Let $\mathcal{G}$ be the subset of $L^2(f_1)$ which contains the elements g such that $\int g(x)f_1(x)dx = 0$ and $\int g^2(x)f_1(x)dx = 1$, and note that $\mathcal{G}$ is compact. Define

$$\sigma^2_{i,m}(g) = \mathrm{E}_i\, g^2(X_1) + 2\sum_{j=1}^{m} \mathrm{E}_i\, g(X_1)g(X_{j+1}) - (2m+1)\left[\mathrm{E}_i\, g(X_1)\right]^2 \tag{2.28}$$

and

$$S_i^{(m)}(g) = \frac{\left[\mu_1(g) - \mu_0(g)\right]^2}{\sigma^2_{i,m}(g)} \tag{2.29}$$

for $i = 0, 1$. If the process is m-dependent under H_i then $\sigma^2_{i,m} = \sigma_i^2$. Finally, let $g_1^{(m)}$

denote the solution to the integral equation (2.19), which by Theorem 3 maximizes $S_1^{(m)}$ over the space $L^2(f_1)$.

Lemma 4. *The functional $S_1^{(m)}$ is continuous.*

Proof. That the numerator of (2.29) is continuous follows from the Schwarz inequality and the assumption (a) of Section 2.3:

$$\left|[\mu_1(g) - \mu_0(g)] - [\mu_1(h) - \mu_0(h)]\right| = |\langle g - h, (f_1 - f_0)/f_1\rangle|$$
$$\leq \|g - h\| \cdot \|(f_1 - f_0)/f_1\|.$$

Thus we must show that $\sigma^2_{1,m}(\cdot)$ is continuous. The first term on the righthand side of (2.28) is the composition of the maps $g \mapsto \|g\|$ and $x \mapsto x^2$, so it is continuous. The map $g \mapsto \int g(x)f_1(x)dx = \langle g, 1\rangle$ is continuous by the Riesz representation theorem. Thus the last term of (2.28) is continuous. Now suppose $(\Omega, \mathcal{F}, P_1)$ is the underlying probability space when H_1 is true, and let $L^2(P_1)$ denote the Hilbert space consisting of all random variables X such that $\mathrm{E}_1 X^2 < \infty$ with the inner product $\langle X, Y\rangle = \mathrm{E}_1 XY$. The random variables $g(X_j)$ for $j = 1, 2, \ldots$ are in $L^2(P_1)$ if g is in $L^2(f_1)$. Also, if $\|\cdot\|_1$ denotes the norm for $L^2(f_1)$ and $\|\cdot\|_2$ denotes the norm in $L^2(P_1)$, then $\|g - h\|_1 = \|g(X_j) - h(X_j)\|_2$ by stationarity. By the continuity of the inner product, then, $|\mathrm{E}_1 g(X_1)g(X_j) - \mathrm{E}_1 h_n(X_1)h_n(X_j)| \to 0$ as $\|g - h_n\|_1 \to 0$. Thus the remaining terms of (2.28) are continuous. □

Define the class $\mathcal{L}$ to consist of all bivariate joint densities which have the diagonal expansion (2.22) with $\{\theta_n\}$ being polynomials.

Lemma 5. *If the densities $f_1^j, j = 1, 2, \ldots$ are in $\mathcal{L}$ then the functional S_1 is continuous.*

Proof. The idea of the proof is to show that $S_1^{(m)}$ converges to S_1 uniformly. First we show that $\sigma_{1,m}^2(g)$ converges uniformly to $\sigma_1^2(g)$ for g in $\mathcal{G}$. Let $\{\theta_n\}$ be the orthonormal functions in the expansion (2.22). Since $g \in L^2(f_1)$, g has an expansion

$$g(x) = \sum_{n=1}^{\infty} b_n \theta_n(x)$$

Then, introducing the expansion (2.22) and applying the orthogonality relation, we have

$$\begin{aligned} &\iint g(x)g(y)f_1^{(j)}(x,y)dxdy \\ &\qquad = \iint \sum_{l=1}^{\infty}\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} b_l b_m a_n^{(j)} \theta_l(x)\theta_n(x)\theta_m(y)\theta_n(y) f_1(x) f_1(y) dxdy \\ &\qquad = \sum_{n=1}^{\infty} b_n^2 a_n^{(j)}. \end{aligned}$$

Since $\int g^2(x) f_1(x)dx = \sum_{n=1}^{\infty} b_n^2 = 1$ for $g \in \mathcal{G}$, the maximum value of $|\iint g(x)g(y)f_1^{(j)}(x,y)dxdy|$ is equal to the maximum value of the sequence $\{|a_n^{(j)}|, n \geq 1\}$. It will be shown in Chapter 4 that this maximum occurs at either $|a_1^{(j)}|$ or $|a_2^{(j)}|$. Since $\theta_i \in \mathcal{G}$ and $\sigma_1^2(\theta_i) = 1 + 2\sum_j a_i^{(j)}$ for $i = 1, 2$, we have $1 + \sum_j c_j < \infty$ where $c_j = \max(|a_1^{(j)}|, |a_2^{(j)}|)$, and this series dominates the series $\sigma_1^2(g)$ independently of g. Thus $\sigma_{1,m}^2(g)$ converges uniformly.

Now if g is any nonconstant element of $L^2(f_1)$, then it follows that there exist constants a, b such that $ag+b \in \mathcal{G}$. Since $\sigma_{1,m}^2(ag+b)$ converges uniformly, and since $S_1^{(m)}$ and S_1 are invariant under such transformations of g, it follows that $S_1^{(m)}(g)$ converges uniformly to $S_1(g)$. □

We are now ready to prove the main result.

Theorem 6. *If the densities $f_1^j, j = 1,2,\ldots$ are in $\mathcal{L}$ then there exists a solution $g_1 \in \mathcal{G}$ to the optimization problem (2.12). If $g_1^{(m)}$ solves the integral equation (2.19), then the sequence $\{S_1(g_1^{(m)})\}$ converges to $S_1(g_1)$ as $m \to \infty$.*

Proof. Since S_1 is continuous and the set $\mathcal{G}$ is compact, there exists an element g' of $\mathcal{G}$ such that S_1 achieves its maximum value on the set $\mathcal{G}$ at g'. If g is any nonconstant element of $L^2(f_1)$, then there exist constants a and b such that $ag + b \in \mathcal{G}$. But $S_1(g) = S_1(ag + b) \leq S_1(g')$. Thus g' solves (2.12).

Let $\epsilon > 0$. In the proof of Lemma 5, it was shown that $S_1^{(m)}$ converges uniformly to S_1. Thus there exists an integer M such that for every $m \geq M$ and every $g \in L^2(f_1)$ we have $|S_1^{(m)}(g) - S_1(g)| < \epsilon$. Let $m \geq M$ be fixed. If $S_1^{(m)}(g_1^{(m)}) < S_1(g_1)$, then we must have $S_1^{(m)}(g_1) \leq S_1^{(m)}(g_1^{(m)}) < S_1(g_1)$. Otherwise, we have $S_1(g_1^{(m)}) \leq S_1(g_1) \leq S_1^{(m)}(g_1^{(m)})$. In either case, $|S_1^{(m)}(g_1^{(m)}) - S_1(g_1)| < \epsilon$. This implies that $|S_1(g_1^{(m)}) - S_1(g_1)| < 2\epsilon$. □

According to the theorem, one can achieve a value of the performance measure S_1 which is arbitrarily close to the optimal value by solving the integral equation (2.19) with m sufficiently large.

CHAPTER 3

THE MINIMAX FORMULATION

3.1 The Performance Measure S_2

Before one can determine the optimal test statistic (or nonlinearity) in some class of allowable test statistics, one must define an ordering on the class. Although the most natural ordering to consider is determined by the receiver operating characteristic (ROC), the ROC itself does not provide a total ordering, so that given two different test statistics one cannot always say which is the better by comparing their ROC's. Rather, in order to have a total ordering, one must specify a particular region, or operating point, of the ROC. Under the Neyman-Pearson formulation, this operating point is specified to be such that $P_0 = \alpha$. Such an operating point may be undesirable when the asymptotic performance is important since only P_1 converges to 0 while P_0 remains fixed. For this reason, it may be better to consider the minimax operating point where $P_0 = P_1$. The term "minimax" refers to the fact that the decision rule, including both the threshold (or the critical region) and the nonlinearity g are chosen to solve the problem

$$\text{minimize} \quad \max_{i=0,1} P_i. \tag{3.1}$$

Actually, we shall consider the problem (3.1) in an asymptotic sense, similar to our method for the Neyman-Pearson formulation. Thus we attempt to maximize the rate at which $\max(P_0, P_1)$ converges to zero. Note that if the ROC is continuous, then we can always choose the critical region so that $P_0 = P_1$, and thus we actually maximize the rate at which the common value of P_0 and P_1 converges to 0. This rate is given approximately by the performance measure S_2, as derived in this section.

We proceed by assuming as in Chapter 2 that the critical region A is given by $[n\gamma_1, n\gamma_2]$, where γ_1 and γ_2 are given by (2.7), so that the error probabilities are given approximately by (2.9). Now if we choose the parameter $\gamma = 0$ so that $v(\gamma) = \mu_1 - \mu_0$, then the expressions for the error probabilities reduce to

$$\begin{aligned} P_0 &= \Phi\left[\sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_0 - \sigma_1}\right] - \Phi\left[\sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_0 + \sigma_1}\right] \\ P_1 &= \Phi\left[-\sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_0 - \sigma_1}\right] + \Phi\left[-\sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_0 + \sigma_1}\right] \end{aligned} \tag{3.2}$$

Now typically the term $\Phi[-\sqrt{n}(\mu_1 - \mu_0)/(\sigma_0 - \sigma_1)]$ is orders of magnitude smaller than the term $\Phi[-\sqrt{n}(\mu_1 - \mu_0)/(\sigma_0 + \sigma_1)]$, and in fact the ratio of these two terms goes to zero:

$$\lim_{n\to\infty} \frac{\Phi\left[-\sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_0 - \sigma_1}\right]}{\Phi\left[-\sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_0 + \sigma_1}\right]} = 0.$$

Thus we may approximate the first term of the expression for P_0 by 1 and approximate the first term of the expression for P_1 by 0. Then we have the error probabilities given approximately by

$$P_0 = P_1 = \Phi\left[-\sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_0 + \sigma_1}\right] \tag{3.3}$$

It should be clear from (3.3) that the quantity $(\mu_1 - \mu_0)/(\sigma_0 + \sigma_1)$ determines (approximately) the rate at which P_0 and P_1 converge to 0. However, if we assume as we did in

Chapter 2 that $\mu_1 > \mu_0$, then this is equivalent to maximizing the quantity

$$S_2 = \frac{(\mu_1 - \mu_0)^2}{(\sigma_0 + \sigma_1)^2}. \tag{3.4}$$

This performance measure S_2 determines approximately the rate at which P_0 and P_1 converge to 0. It has been derived also in [3] using Chernoff bounds as a nonlocal approach to the signal detection problem (1.11). The approach here is different in that the focus is on signal discrimination. Note that the performance measure S_2 treats equally the asymptotic variance under the two hypotheses, whereas this is not the case for the performance measure S_1. However, in the remainder of this chapter we shall see that S_2 is much more difficult to work with analytically due to the nonlinear function of σ_0^2 and σ_1^2 in the denominator.

3.2 The Optimal Nonlinearity

As in the case of the performance measure S_1 of the last chapter, the nonlinearity g_2 which maximizes the performance measure S_2 is also given by the solution of an integral equation; however, the integral equation for this case is nonlinear, and we shall not be able to obtain a closed form solution. The technique used to derive the integral equation is similar to that used earlier, and we will again assume that the processes are m-dependent. The value of $S_2(g)$ remains unchanged if g is multiplied by a constant, and we may therefore attempt to maximize $(\mu_1 - \mu_0)^2$ with $(\sigma_0 + \sigma_1)^2$ constrained to be constant. Equivalently, we can consider the following optimization problem which involves a Lagrange multiplier λ:

$$\text{maximize} \quad \mu_1 - \mu_0 - \lambda(\sigma_0 + \sigma_1)^2 \tag{3.5}$$

To solve this problem, define $J(g) = \mu_1(g) - \mu_0(g) - \lambda\left[\sigma_0(g) + \sigma_1(g)\right]^2$, and let δg be an arbitrary continuous function satisfying $\mathrm{E}_i \delta g^2(X_1) < \infty$ for $i = 0, 1$. A necessary condition for g to solve (3.5) is that $\frac{d}{d\epsilon} J(g + \epsilon\delta g)\big|_{\epsilon=0} = 0$. To show the steps involved in taking the derivative, write $J(g)$ in the following form:

$$J(g) = \mu_1(g) - \mu_0(g) - \lambda\sigma_0^2(g) - \lambda\sigma_1^2(g) - 2\lambda\sqrt{\sigma_0^2(g)\sigma_1^2(g)} \tag{3.6}$$

Also define $A_i(g, \delta g)$ by $2A_i(g, \delta g) = \frac{d}{d\epsilon}\sigma_i^2(g + \epsilon\delta g)\big|_{\epsilon=0}$. Then $A_i(g, \delta g)$ is given by

$$\begin{aligned} A_i(g, \delta g) = {} & \mathrm{E}_i g(X_1)\delta g(X_1) + \sum_{j=1}^{m}\left[\mathrm{E}_i g(X_1)\delta g(X_j) + \mathrm{E}_i g(X_j)\delta g(X_1)\right] \\ & -(2m+1)\mathrm{E}_i g(X_1)\mathrm{E}_i \delta g(X_1) \end{aligned} \tag{3.7}$$

Now we have for the contribution of the last term on the right of (3.6)

$$\begin{aligned} \frac{d}{d\epsilon}\sqrt{\sigma_0^2(g + \epsilon\delta g)\sigma_1^2(g + \epsilon\delta g)}\Big|_{\epsilon=0} &= \left[\sigma_0^2(g)\sigma_1^2(g)\right]^{-\frac{1}{2}}\left[\sigma_0^2(g)A_1(g, \delta g) + A_0(g, \delta g)\sigma_1^2(g)\right] \\ &= \frac{\sigma_0(g)}{\sigma_1(g)}A_1(g, \delta g) + \frac{\sigma_1(g)}{\sigma_0(g)}A_0(g, \delta g) \end{aligned} \tag{3.8}$$

The contributions of the other terms are immediate. Thus we have

$$\begin{aligned} & \frac{d}{d\epsilon} J(g + \epsilon\delta g)\Big|_{\epsilon=0} \\ & \quad = \mathrm{E}_1 \delta g(X_0) - \mathrm{E}_0 \delta g(X_0) - 2\lambda\left[\left(1 + \frac{\sigma_1}{\sigma_0}\right)A_0(g, \delta g) + \left(1 + \frac{\sigma_0}{\sigma_1}\right)A_1(g, \delta g)\right] \end{aligned} \tag{3.9}$$

We obtain the integral equation by introducing the densities f_0, f_1, $f_0^j, j = 1 \ldots m$, and $f_1^j, j = 1 \ldots m$ into (3.9) and setting the result equal to 0. This yields

$$\begin{aligned} 0 = \int \delta g(x)\Big\{ & f_1(x) - f_0(x) - 2\lambda\left(1 + \frac{\sigma_1}{\sigma_0}\right)\left[g(x)f_0(x) + \int \tilde{K}_0(x, y)g(y)dy\right] \\ & - 2\lambda\left(1 + \frac{\sigma_0}{\sigma_1}\right)\left[g(x)f_1(x) + \int \tilde{K}_1(x, y)g(y)dy\right]\Big\}dx \end{aligned} \tag{3.10}$$

where $\tilde{K}_i$ is given by

$$\tilde{K}_i(x,y) = \sum_{j=1}^{m} [f_i^j(x,y) + f_i^j(y,x)] - (2m+1) f_i(x) f_i(y). \tag{3.11}$$

The equation (3.10) holds for arbitrary δg iff the expression in braces is identically zero. Thus g must solve the integral equation

$$2\lambda g(x) = \frac{f_1(x) - f_0(x)}{(1+\tau) f_0(x) + (1+\tau^{-1}) f_1(x)} - 2\lambda \int L(x,y) g(y) dy \tag{3.12}$$

where $\tau = (\sigma_1/\sigma_0)$ and the kernel L is given by

$$L(x,y) = \frac{(1+\tau)\tilde{K}_0(x,y) + (1+\tau^{-1})\tilde{K}_1(x,y)}{(1+\tau) f_0(x) + (1+\tau^{-1}) f_1(x)}. \tag{3.13}$$

Again we observe that λ determines the scaling of g. Thus the particular value of λ is not significant except that it must have the proper sign so that if g solves the integral equation (3.12), then $\mu_1(g) > \mu_0(g)$. Consider now

$$\begin{aligned} [\sigma_0(g) + \sigma_1(g)]^2 &= (1+\tau)\sigma_0^2(g) + (1+\tau^{-1})\sigma_1^2(g) \\ &= \int g(x) \Big\{ \big[(1+\tau) f_0(x) + (1+\tau^{-1}) f_1(x)\big] g(x) \\ &\qquad \int \big[(1+\tau)\tilde{K}_0(x,y) + (1+\tau^{-1})\tilde{K}_1(x,y)\big] g(y) dy \Big\} dx \end{aligned} \tag{3.14}$$

If g solves the integral equation (3.12) for $\lambda = \frac{1}{2}$, then the expression in the braces in (3.14) reduces to $f_1(x) - f_0(x)$, and thus $[\sigma_0(g) + \sigma_1(g)]^2 = \mu_1(g) - \mu_0(g) > 0$. We shall therefore assign to λ the value $\frac{1}{2}$ and henceforth consider the integral equation

$$g(x) = \frac{f_1(x) - f_0(x)}{(1+\tau) f_0(x) + (1+\tau^{-1}) f_1(x)} - \int L(x,y) g(y) dy. \tag{3.15}$$

Furthermore, we observe that if g_2 solves the integral equation (3.15) then $S_2(g_2) = \mu_1(g_2) - \mu_0(g_2)$, a result which is similar to the one obtained in Chapter 2 for the optimal value of S_1. We have proved the following theorem.

Theorem 7. *If the process $\{X_i\}$ is m-dependent under both H_0 and H_1, then a sufficient condition for g_2 to maximize S_2 is that g_2 solve the integral equation (3.15). Furthermore, if g_2 solves (3.15) then $S_2(g_2) = \mu_1(g_2) - \mu_0(g_2)$.*

In comparing the integral equation (3.15) with the integral equation (2.19), we note first of all that (3.15) is nonlinear because of the fact that τ is a function of g. Let us consider now what happens when τ varies. If τ is very small, then σ_1 is much smaller than σ_0, and thus the value of performance measure S_2 is very close to that of the performance measure S_0. In fact, $S_2 \to S_0$ as $\tau \to 0$. Now observe that the integral equation (3.15), when rescaled, converges as $\tau \to 0$ to the integral equation (2.19) which maximizes the performance measure S_1. This provides us with some insight to the relation between the performance measures S_0, S_1, and S_2 and the role that τ plays in the integral equation (3.15). We observe, for example, that there is a conflict of objectives for very small τ in that the value of the performance measure S_2 is approximately equal to that of S_0, while the integral equation (3.15) provides a nonlinearity which is close to the one which maximizes the performance measure S_1. A similar conflict occurs as τ approaches ∞, with the roles of S_0 and S_1 reversed. Of course, there is no conflict of objective if $S_0 \approx S_1$, but this implies that $\tau \approx 1$. Thus we expect that τ will have a "reasonable" value on the order of one. We find this to be the case in Chapter 5 where a numerical solution to the integral equation (3.15) is found.

3.3 The Solution of the Integral Equation

The equation (3.15) is nonlinear because τ is a function of g, and for this reason, finding a closed form solution is rather difficult. If, however, we had clairvoyance to

know the correct value of τ, then we could find the solution g_2 by solving a linear integral equation. In fact, we might try to guess the value of τ, find the solution of the resulting linear integral equation, and then compute τ to verify if our guess was correct. This suggests an iterative method where the computed value of τ from the previous solution becomes the new value for τ at the next iteration of the procedure. This method is used to obtain a numerical solution to (3.15) in Chapter 5; it is found that the successive values of τ do in fact converge.

Although we cannot find a closed form solution to (3.15), we may treat τ as a constant whose value is unknown, and thereby extend the analysis relating to the equation (3.15). If we make the substution

$$h(x) = g(x)\sqrt{(1+\tau)f_0(x) + (1+\tau^{-1})f_1(x)},$$

we obtain the integral equation

$$h(x) = \frac{f_1(x) - f_0(x)}{\sqrt{w_\tau(x)}} + \int L^*(x,y)h(y)dy \tag{3.16}$$

where the symmetric kernel L^* is given by

$$L^*(x,y) = \frac{(1+\tau)\tilde{K}_0(x,y) + (1+\tau^{-1})\tilde{K}_1(x,y)}{\sqrt{w_\tau(x)w_\tau(y)}} \tag{3.17}$$

and where w_τ is defined by

$$w_\tau(t) = (1+\tau)f_0(t) + (1+\tau^{-1})f_1(t). \tag{3.18}$$

For a given value of τ, the integral equation (3.16) is a Fredholm equation of the second kind, provided we have the conditions

$$(a) \quad \int \frac{[f_1(x) - f_0(x)]^2}{w_\tau(x)}dx < \infty$$

$$(b) \quad \iint |L^*(x,y)|^2 dxdy < \infty.$$

These conditions imply that the solution h is square integrable, and then it follows that that $E_i g^2(X_1) < \infty$ for $i = 0, 1$. Note that we do not require the condition that $f_0(x)/f_1(x)$ be bounded as we did for the integral equation (2.19). Condition (a) follows from the fact that $|f_1(x) - f_0(x)|/w_\tau(x)$ is bounded by $(1+\tau)^{-1} + (1+\tau^{-1})^{-1}$. To show that condition (b) holds, it suffices to show that

$$\iint \left| \frac{\tilde{K}_i(x,y)}{\sqrt{w_\tau(x) w_\tau(y)}} \right|^2 dx dy < \infty \tag{3.19}$$

since we may then apply the Minkowski inequality. If all the joint densities involved have the expansion (2.22), then the inequality (3.19) follows from a similar argument for the case of the kernel K_1^* in Chapter 2. For example, consider the terms of the form

$$\frac{f_0^j(x,y) f_0^k(x,y)}{w_\tau(x) w_\tau(y)} \leq \frac{f_0^j(x,y) f_0^k(x,y)}{(1+\tau)^2 f_0(x) f_0(y)}.$$

It was shown in Chapter 2 that such terms are integrable. Thus condition (b) holds. Since the integral equation (3.16) has a symmetric kernel, the Hilbert-Schmidt theory applies as in Chapter 2. If the eigenvalues and eigenvectors of the kernel (3.17) are denoted by $\{\lambda_n\}$ and $\{\phi_n\}$, then a solution h_2 of the integral equation (3.16) has the expansion

$$\begin{aligned} h_2(x) &= h^*(x) + \sum_{n=0}^{\infty} \frac{\lambda_n c_n}{1-\lambda_n} \phi_n(x) \\ &= \sum_{n=0}^{\infty} \frac{c_n}{1-\lambda_n} \phi_n(x) \end{aligned} \tag{3.20}$$

where

$$h^*(x) = \frac{f_1(x) - f_0(x)}{\sqrt{w_\tau(x)}}$$

and

$$c_n = \int h^*(x) \phi_n(x) dx.$$

The solution is unique if and only if $\lambda_n \neq 1$ for all n. Since we do not have clairvoyance to know the true value of τ, the solution (3.20) is purely academic.

If the process is iid, then the kernel L from (3.13) has the simpler form

$$L(x,y) = -\frac{(1+\tau)f_0(x)f_0(y) + (1+\tau^{-1})f_0(x)f_1(y)}{(1+\tau)f_0(x) + (1+\tau^{-1})f_1(x)}$$

and the integral equation (3.15) has the solution

$$g_{iid}(x) = \frac{B_0 f_0(x) + B_1 f_1(x)}{(1+\tau)f_0(x) + (1+\tau^{-1})f_1(x)} \tag{3.21}$$

where $B_0 = [(1+\tau)\mu_0 - 1]$ and $B_1 = [(1+\tau^{-1})\mu_1 + 1]$. There are three unknown quantities in the expression (3.21): τ, μ_0, and μ_1. These quantities can be found by solving the following system of nonlinear equations:

$$\begin{aligned} \mu_0 &= \int g_{iid}(x) f_0(x) dx \\ \mu_1 &= \int g_{iid}(x) f_1(x) dx \\ \tau &= \sqrt{\frac{\sigma_1^2(g_{iid})}{\sigma_0^2(g_{iid})}}. \end{aligned} \tag{3.22}$$

Note that $g_{iid}(x) + C$ solves (3.15) for an arbitrary constant C. We may therefore take an arbitrary value for either μ_0 or μ_1. In fact, with τ fixed the first two equations in (3.22) are linear in μ_0 and μ_1 and are singular. Therefore the system (3.22) does not have a unique solution.

3.4 The Performance Measure S_3

The performance measures S_0 and S_1 which were derived in Chapter 2 have the undesirable feature that they treat unequally the performance under the two hypotheses,

and yet they are relatively nice regarding analysis since they lead to linear integral equations. On the other hand, the performance measure S_2 treats both hypotheses equally but leads to a nonlinear integral equation. We are led to consider also the performance measure

$$S_3 = \frac{[\mu_1 - \mu_0]^2}{\sigma_0^2 + \sigma_1^2} \tag{3.23}$$

which treats both hypotheses equally and leads to a linear integral equation as well. This performance measure is derived in this section by considering Chernoff bounds for the error probabilities, and is actually a slight modification of the method of Sadowsky and Bucklew [3] in which they derived the performance measure S_2.

If the test statistic T_n has a normal distribution with mean $n\mu_i$ and variance $n\sigma_i^2$ then these are the Chernoff bounds:

$$\begin{aligned} P[T_n \geq n\gamma] &\leq \exp[-nI_i(\gamma)] \qquad \text{if } \mu_i < \gamma \\ P[T_n \leq n\gamma] &\leq \exp[-nI_i(\gamma)] \qquad \text{if } \mu_i > \gamma \end{aligned} \tag{3.24}$$

where

$$I_i(\gamma) = \frac{(\mu_i - \gamma)^2}{2\sigma_i^2}. \tag{3.25}$$

The Chernoff bounds are asymptotically tight in the sense that

$$\begin{aligned} \lim_{n\to\infty} -\frac{1}{n} \log P[T_n \geq n\gamma] &= I_i(\gamma) \qquad \text{if } \mu_i < \gamma \\ \lim_{n\to\infty} -\frac{1}{n} \log P[T_n \leq n\gamma] &= I_i(\gamma) \qquad \text{if } \mu_i > \gamma \end{aligned}$$

and they hence can provide good approximations to the error probabilites if n is large. Of course, if the distribution of T_n is only approximately normal, then the bounds given by (3.24) are only approximations of the true Chernoff bounds. Nevertheless, we shall proceed under the assumption that such approximations are acceptable. From (3.24) we

see that if $n\gamma$ is the threshold and $\mu_0 < \gamma < \mu_1$, then I_i determines (approximately) the bound for the error probability P_i. Since a larger value for I_i results in a smaller bound for P_i, it is desirable to make both I_0 and I_1 as large as possible. It is obvious that I_i is a convex function of γ which takes its minimum value at $\gamma = \mu_i$. Thus as γ increases from μ_0 to μ_1, I_0 increases and I_1 decreases. Sadowsky and Bucklew [3] proceeded from this point by maximizing the $\min(I_0, I_1)$ and obtained the result that

$$S_2 = \max_{\mu_0<\gamma<\mu_1} \min_{i=0,1} I_i(\gamma).$$

It is fairly straightforward, however, to show that

$$S_3 = \max_{\mu_0<\gamma<\mu_1} [I_0(\gamma) + I_1(\gamma)]$$

and this justifies the use of performance measure S_3.

The following method for maximizing S_3 is very similar to the method in Chapter 2 for maximizing S_1, and we must assume that the processes are m-dependent. We know that g maximizes S_3 if and only if g maximizes $J_3(g) = \mu_1(g) - \mu_0(g) - \lambda[\sigma_0^2(g) + \sigma_1^2(g)]$, where λ is a Lagrange multiplier. The condition that $J(\cdot)$ vanish at g leads to the integral equation

$$2\lambda g(x) = \frac{f_1(x) - f_0(x)}{f_0(x) + f_1(x)} - 2\lambda \int M(x,y)g(y)dy \tag{3.26}$$

where the kernel M is given by

$$M(x,y) = \frac{\tilde{K}_0(x,y) + \tilde{K}_1(x,y)}{f_0(x) + f_1(x)}.$$

The functional J_3 is convex in g, provided $\lambda > 0$, so that the condition that g solve the integral equation (3.26) is also a sufficient condition for g to maximize S_3. We therefore take $\lambda = \frac{1}{2}$ and the integral equation (3.26) becomes

$$g(x) = \frac{f_1(x) - f_0(x)}{f_0(x) + f_1(x)} - \int M(x,y)g(y)dy. \tag{3.27}$$

We can also show easily that if g_3 solves the integral equation (3.27), then $\sigma_0^2(g_3) + \sigma_1^2(g_3) = \mu_1(g_3) - \mu_0(g_3)$ so that the optimal value of S_3 is $S_3(g_3) = \mu_1(g_3) - \mu_0(g_3)$. These results are summarized in Theorem 8.

Theorem 8. *If the process $\{X_i\}$ is m-dependent under both H_0 and H_1, then a sufficient condition for g_3 to maximize S_3 is that g_3 solve the integral equation (3.27). Furthermore, if g_3 solves (3.27) then $S_3(g_3) = \mu_1(g_3) - \mu_0(g_3)$.*

The integral equation (3.27) can be transformed into an integral equation with a symmetric kernel by making the substitution $h(x) = g(x)\sqrt{f_0(x) + f_1(x)}$ so that the Hilbert-Schmidt theory applies as before. We shall not pursue this further. We shall, however, proceed to find the optimal nonlinearity for iid processes. If the processes are both iid, then the kernel M has the form

$$M(x,y) = -\frac{f_0(x)f_0(y) + f_1(x)f_1(y)}{f_0(x) + f_1(x)}$$

and the integral equation gives us immediately the form of the iid solution:

$$g_{iid}(x) = \frac{B_0 f_0(x) + B_1 f_1(x)}{f_0(x) + f_1(x)} \tag{3.28}$$

where $B_0 = \mu_0 - 1$ and $B_1 = \mu_1 + 1$. To find the unknown constants B_0, B_1, we subtitute for g_{iid} in the linear equations

$$\begin{aligned} \mu_0 = B_0 + 1 &= \int g_{iid}(x) f_0(x) dx \\ \mu_1 = B_1 - 1 &= \int g_{iid}(x) f_1(x) dx. \end{aligned} \tag{3.29}$$

The system (3.29) is in fact singular, so that we may take either B_0 or B_1 to be arbitrary. Therefore we shall arbitrarily take $B_0 = 0$ and this gives us the value

$$B_1 = \left[\int \frac{f_0(x)f_1(x)}{f_0(x) + f_1(x)} dx\right]^{-1}. \tag{3.30}$$

Thus the iid solution has been determined explicitly.

3.5 Extension to ϕ-Mixing Processes

The results proved in Section 2.4 are also true for the performance measures S_2 and S_3. Because of the similarity of these two performance measures, the proof for either case is nearly identical; therefore only the results for S_2 will be stated and proved. The notation is as follows: Define the density $w = \frac{1}{2}(f_0 + f_1)$ and let $L^2(w)$ be the Hilbert space of Borel functions g such that $\int g^2(x)w(x)dx < \infty$ with the inner product $\langle g, h \rangle = \int g(x)h(x)w(x)dx$. Note that $g \in L^2(w)$ implies that $\mathrm{E}_i\, g^2(X_1) < \infty$ for $i = 0, 1$. Let $\sigma^2_{i,m}(g)$ be defined by (2.28) and define

$$S_2^{(m)}(g) = \frac{\left[\mu_1(g) - \mu_0(g)\right]^2}{\left[\sigma_{0,m}(g) + \sigma_{1,m}(g)\right]^2}. \tag{3.31}$$

We denote by $\mathcal{L}$ the class of joint densities which have a diagonal expansion, as in Section 2.4.

Lemma 9. *The functional $S_2^{(m)}$ is continuous.*

Proof. It was shown in the proof of Lemma 4 that the numerator of (3.31) is continuous and that $\sigma^2_{i,m}(g)$ is continuous. Since additions, square roots, divisions, etc. preserve continuity, it follows that $S_2^{(m)}$ is continuous. □

Lemma 10. *If the joint densities f_i^j, $i = 0, 1$, $j = 1, 2, \ldots$, are in $\mathcal{L}$, then the functional S_2 is continuous.*

Proof. The proof consists of showing that $S_2^{(m)}(g)$ converges to $S_2(g)$ uniformly for $g \in L^2(w)$. Define $\mathcal{G}_i$ to be the class of all functions $g \in L^2(w)$ such that $\mathrm{E}_i\, g(X_1) = 0$ and

$E_i g^2(X_1) = 1$. From the proof of Lemma 5, it follows that $\sigma^2_{i,m}(g)$ converges uniformly for $g \in \mathcal{G}_i$. Now suppose that g is an arbitrary nonconstant element of $L^2(w)$. Then there exist constants a_i and b_i such that $g_i = a_i g + b_i \in \mathcal{G}_i$, and obviously $\sigma^2_{i,m}(g_i) = a_i^2 \sigma^2_{i,m}(g)$ for all m and $\sigma_i^2(g_i) = a_i^2 \sigma_i^2(g)$. Let $c = |a_1/a_0|$. Then we can write

$$\left|S_2(g) - S_2^{(m)}(g)\right| = \left|\left[\frac{\mu_1(g_1) - \mu_0(g_1)}{c\sigma_0(g_0) + \sigma_1(g_1)}\right]^2 - \left[\frac{\mu_1(g_1) - \mu_0(g_1)}{c[\sigma_0(g_0) + \epsilon_0] + [\sigma_1(g_1) + \epsilon_1]}\right]^2\right| \quad (3.32)$$

where ϵ_0 and ϵ_1 converge to 0 uniformly for $g \in L^2(w)$ as $m \to \infty$. The righthand side of (3.32) is continuous as a function of c for $c \in [0, \infty)$ and approaches 0 as c approaches ∞. Thus for ϵ_0 and ϵ_1 fixed, the righthand side attains its maximum as c varies, and this maximum converges to 0 as ϵ_0 and ϵ_1 approach 0. Hence convergence of $S_2^{(m)}$ is uniform. □

We are now ready to prove the main result.

Theorem 11. *If the joint densities f_i^j, $i = 0, 1$, $j = 1, 2, \ldots$, are in $\mathcal{L}$, then there exists a function $g_2 \in L^2(w)$ which maximizes S_2. If $g_2^{(m)}$ solves the integral equation (3.15), then the sequence $\{S_2(g_2^{(m)})\}$ converges to $S_2(g_2)$ as $m \to \infty$.*

Proof. Define $\mathcal{G}'$ to be the subset of $L^2(w)$ consisting of the vectors g satisfying $\int g(x)^2 w(x) dx = 1$. Since S_2 is continuous and the set $\mathcal{G}'$ is compact, there exists an element g_2 of $\mathcal{G}'$ such that S_2 achieves its maximum value on the set $\mathcal{G}'$ at g_2. If g is any nonconstant element of $L^2(w)$, then there exist constants a and b such that $ag + b \in \mathcal{G}'$. But $S_2(g) = S_2(ag + b) \leq S_2(g_2)$. Thus g_2 maximizes S_2

Let $\epsilon > 0$. By the proof of Lemma 10, $S_2^{(m)}$ converges uniformly to S_2. Thus there exists an integer M such that for every $m \geq M$ and every $g \in L^2(w)$ we have $|S_2^{(m)}(g) - S_2(g)| < \epsilon$. Let $m \geq M$ be fixed. If $S_2^{(m)}(g_2^{(m)}) < S_2(g_2)$, then we must have

$S_2^{(m)}(g_2) \leq S_2^{(m)}(g_2^{(m)}) < S_2(g_2)$. Otherwise, we have $S_2(g_2^{(m)}) \leq S_2(g_2) \leq S_2^{(m)}(g_2^{(m)})$. In either case, $|S_2^{(m)}(g_2^{(m)}) - S_2(g_2)| < \epsilon$, and this implies that $|S_2(g_2^{(m)}) - S_2(g_2)| < 2\epsilon$.

□

CHAPTER 4

MINIMAX ROBUSTNESS

4.1 The Robustness Problem

When the actual probability distributions of the observed processes are precisely the same as the distributions which were assumed in deriving a particular test, then this is referred to as a *matched* situation. The performance of a test in the matched situation is certainly an important consideration. However, also of great importance is the performance of the test under the *mismatched* situation, where the actual probability distributions are close to but slightly different from the assumed distributions. If a given decision rule performs relatively well in the mismatched situation, as compared to the matched situation, then such a decision rule is said to be *robust*. It is the purpose of this chapter to address the issue of robustness in relation to the performance measures S_1, S_3 and the corresponding optimal test statistics as given in the preceding chapters.

To begin, we must first make more precise mathematically the discussion of the preceding paragraph. One approach to robustness which has become very popular and which will be considered here is minimax robustness, a game theoretic approach. Define the $\mathcal{Q}_0$ and $\mathcal{Q}_1$ to be classes of distributions which are possible under the hypotheses

H_0 and H_1, respectively. These *uncertainty classes* are to contain the "nominal" distributions (those distributions which are assumed initially) as well as those distributions which are only slightly different from the nominal distributions. Now define the *least favorable* distributions $F_0^* \in \mathcal{Q}_0$ and $F_1^* \in \mathcal{Q}_1$ to be the distibutions which together with the nonlinearity g^* form a saddle point for the performance measure S as follows:

$$S(g, F_0^*, F_1^*) \leq S(g^*, F_0^*, F_1^*) \leq S(g^*, F_0, F_1) \tag{4.1}$$

where g is any other allowable nonlinearity and F_0 and F_1 are arbitrary distributions from the classes $\mathcal{Q}_0$ and $\mathcal{Q}_1$. In this case g^* is called a minimax robust nonlinearity, and has the property that for any pair of distributions in $\mathcal{Q}_0$ and $\mathcal{Q}_1$, the value of S evaluated at g^* is guaranteed to be at least $S(g^*, F_0^*, F_1^*)$.

The idea of this approach is like that of a game in which nature chooses the distributions F_0, F_1 out of the classes $\mathcal{Q}_0$, $\mathcal{Q}_1$ and the human player chooses the nonlinearity g, the performance measure $S(g, F_0, F_1)$ being the payoff. The first inequality in (4.1) is usually not difficult to show. Indeed, if the least favorable distributions are known then finding the nonlinearity g^* is merely the problem considered in Chapter 2 or Chapter 3. What is usually more difficult is finding the least favorable distributions F_0^*, F_1^* and showing the second inequality in (4.1). Obviously g^*, F_0^*, and F_1^* solve the minimax problem

$$\min_{F_0, F_1} \max_{g} S(g, F_0, F_1) \tag{4.2}$$

and in fact solving (4.2) is the simplest way to find F_0^* and F_1^*, if they exist. That the solution g^*, F_0^*, F_1^* of (4.2) satisfies the left inequality in (4.1) is obvious, and the main task in proving a result in robustness is showing that such a solution also satisfies the right inequality. Equivalently, one might try to show that g^*, F_0^*, F_1^* solve the maximin

problem

$$\max_{g} \min_{F_0, F_1} S(g, F_0, F_1) \tag{4.3}$$

since the solution of (4.3) satisfies the right inequality in (4.1).

If one defines a metric on some class $\mathcal{M}$ of probability distributions, then a quite natural way to define an uncertainty class $\mathcal{Q}$ about a nominal distribution $\tilde{F}$ is to include all distributions in $\mathcal{M}$ which are at a distance ϵ or less from $\tilde{F}$. Such a definition might be appropriate for minimax robustness if one wishes to exploit continuity properties of a performance measure, since by continuity there will be a small change in the performance if the distance between the distributions is small. The ϵ-contamination class which we shall consider here is useful in a different sense and is defined by $\mathcal{Q} = \{F : F = (1-\epsilon)\tilde{F} + \epsilon H, H \in \mathcal{M}\}$. Evidently, every distribution in the class $\mathcal{Q}$ is a mixture of the nominal distribution $\tilde{F}$ and some unknown distribution H with weights $(1-\epsilon)$ and ϵ. The corresponding physical interpretation is that an observation comes from the distribution $\tilde{F}$ with probability $(1-\epsilon)$ and from the distribution H with probability ϵ. Thus if F is a univariate distribution and the process is iid, then out of n observations, approximately $(1-\epsilon)n$ will be from the distribution $\tilde{F}$ and approximately ϵn will be "corrupted" observations from the distribution H. Therefore in the iid case such an uncertainty class has a pleasing physical interpretation. However, if the process is not iid, as we wish to assume, then F is an n dimensional distribution and the interpretation is that with probability ϵ the distribution is completely unknown. This interpretation is not particularly desirable, and we shall therefore modify this particular uncertainty model.

Since the performance measures we have derived involve only the marginal and bivariate joint densities, we shall attempt to define uncertainty classes which involve

only these densities. In particular, we shall assume that the nominal distribution for our uncertainty class $\mathcal{Q}$ is iid with marginal density $\tilde{f}$. The class $\mathcal{Q}$ is then defined to contain all stationary process distributions F such that the marginal density corresponding to F is contained in an ϵ-contamination class about the nominal $\tilde{f}$, and such that the bivariate joint distributions satisfy the condition

$$\sup_g \frac{|\mathrm{Cov}[g(X_1), g(X_{j+1})]|}{\sqrt{\mathrm{Var}\, g(X_1)\, \mathrm{Var}\, g(X_{j+1})}} \leq r_j \tag{4.4}$$

where g ranges over all measurable functions satisfying $\mathrm{E}\, g^2(X_1) < \infty$. Since we assume stationarity, the denominator is actually just $\mathrm{Var}\, g(X_1)$. The condition (4.4) is our way of allowing for some uncertainty in the dependency structure of the process, so that not every process in the class $\mathcal{Q}$ will be iid. Thus the uncertainty class $\mathcal{Q}$ is specified by giving the nominal density $\tilde{f}$, the parameter ϵ, and the sequence $\{r_i\}$.

In the analysis that follows, the least favorable marginal and bivariate joint distributions are derived, and two issues regarding these least favorable distributions must be addressed. First, it must be shown that there do in fact exist stochastic processes having the prescribed distributions, and second, it must be shown that the processes satisfy a mixing condition, so that the central limit theory may be applied as in the preceding chapters. The necessary results for these two issues have been derived by Sadowsky [9] and we shall adapt them as needed. For a fixed marginal distribution function F, the least favorable bivariate distribution functions are given by (F^j being the joint distribution function for X_1 and X_{j+1})

$$F^j(x, y) = (1 - r_j)F(x)F(y) + r_j F(x \wedge y), \quad j = 1, 2, \ldots \tag{4.5a}$$

where $x \wedge y$ is the minimum of x and y. If the distribution function F has a density f,

then we may write for the bivariate densities

$$f^j(x,y) = (1-r_j)f(x)f(y) + r_j\delta(x-y)f(x), \quad j = 1,2,\ldots. \tag{4.5b}$$

It can easily be seen that for such distributions equality is achieved in (4.4) for any g such that $\mathrm{E}\,g^2(X_1) < \infty$. When $\sigma^2(g)$ is computed using the bivariate distributions given by (4.5) and $R = \sum r_j < \infty$, then we have

$$\sigma^2(g) = (1+2R)\mathrm{Var}\,g(X_1) \tag{4.6}$$

which depends only on the marginal density. Furthermore, if we define $\mathcal{G}$ to be the class of all measurable functions g such that $\mathrm{Var}\,g(X_1) = 1$, then the supremum of σ^2 over $\mathcal{G}$ is achieved by every g in $\mathcal{G}$ and is equal to $(1+2R)$. That a process exists which has distributions given by (4.5) is shown in two different constructive proofs by Sadowsky [9]. Let $\{\theta_i\}$ be a sequence of nonnegative real numbers such that $\sum_{i=1}^{\infty} \theta_i = 1$. Then in these two constructions the r-sequences are given by $r_j = \sum_{m=0}^{\infty} \theta_m\theta_{j+m}$ in one case and by $r_j = \sum_{m=j+1}^{\infty} \frac{m-j}{m}\theta_m$ in the other. Thus in Sadowsky's constructions, arbitrary r-sequences may not be possible. Note that if the sequence $\{\theta_i\}$ takes positive values only for $i = 1,2,\ldots,m$, then the processes in either construction are m-dependent.

The following central limit theorem, from [9], is appropriate:

Theorem 12. *Let $0 < \delta \le \infty$ and set $q = \delta/(2+\delta)$ if $\delta < \infty$ or set $q = 1$ if $\delta = \infty$. Let $\{X_i\}$ be a stationary strong mixing process which satisfies (1.9) with*

$$\sum_{j=1}^{\infty} \alpha_j^q < \infty, \tag{4.7}$$

and let g be a Borel function such that $\mathrm{E}\,|g(X_1)|^{2+\delta} < \infty$ if $\delta < \infty$ or such that $|g(X_1)|$ is almost surely bounded if $\delta = \infty$. Then the sum $\sigma^2(g)$ defined in (1.8) converges, and $[T_n(\mathbf{X}) - n\mu(g)]/\sqrt{n\sigma^2(g)}$ converges in distribution to a standard normal random variable, provided $\sigma^2(g) > 0$.

Further results from [9] show that the process is strong mixing and the condition (4.7) is satisfied if there exist constants $K > 0$ and $\epsilon > 0$ such that $r_j \leq Kj^{-(1+2q+\epsilon)/q}$. Thus if the r-sequence is dominated by an exponential sequence, the condition (4.7) holds. To apply the theorem, then, it remains only to show that $\mathrm{E}|g(X_1)|^{2+\delta} < \infty$, and this holds in particular if g is bounded. Thus the two issues mentioned above are resolved.

Because the condition (4.4) is defined in terms of a supremum over second-order functions g, it is not directly obvious whether a given bivariate distribution satisfies such a bound. In order to determine whether the bound holds for a given bivariate density f^j, it is useful to consider the diagonal diagonal expansion (2.22), if it exists. Any g which satisfies $\int g^2(x)f(x)dx < \infty$ has an expansion $g(x) = \sum_n b_n\theta_n(x)$, and for such an expansion, we have as well that

$$\int g^2(x)f(x)dx = \sum_n b_n^2$$

$$\iint g(x)g(y)f^j(x,y)dxdy = \sum_n b_n^2 a_n^{(j)}$$

$$\int g(x)f(x)dx = b_0.$$

These expressions imply that

$$\frac{|\mathrm{Cov}[g(X_1), g(X_{j+1})]|}{\mathrm{Var}\, g(X_1)} = \frac{\left|\sum_{n\geq 1} b_n^2 a_n^{(j)}\right|}{\sum_{n\geq 1} b_n^2} \tag{4.8}$$

Since $\mathrm{Cov}[g(X_1), g(X_{j+1})]/\mathrm{Var}\, g(X_1)$ is invariant under the scaling of g, we can assume without loss of generality that $\sum_n b_n^2 = 1$, so that the denominator of (4.8) is equal to 1. Then it is obvious that the sup in (4.4) is obtained by θ_i where i is such that $|a_i^{(j)}| = \max\{|a_n^{(j)}|, n \geq 1\}$. If the orthonormal functions $\{\theta_n\}$ are polynomials (that is, f^j is in the class $\mathcal{L}$ defined in Chapter 2) then this maximum coefficient occurs as either $a_1^{(j)}$ or $a_2^{(j)}$. To show this, we require a fact from [12] that for any such diagonal expansion in which the orthonormal functions are polynomials, there exists a probability density function h_j having support in the interval $[-1, 1]$ such that $a_n^{(j)} = \int_{-1}^{1} t^n h_j(t)dt$. Then for $n > 2$ it is obvious that

$$|a_n^{(j)}| \leq \int_{-1}^{1} |t|^n h_j(t)dt \leq \int_{-1}^{1} t^2 h_j(t)dt = |a_2^{(j)}|.$$

so that the assertion holds. Let $\xi_n = \int x^n f(x)dx$, $\zeta_{mn}^{(j)} = \iint x^m y^n f^j(x,y)dxdy$, and $\sigma^2 = \xi_2 - \xi_1^2$. Then for this case $a_1^{(j)}$ and $a_2^{(j)}$ are given by

$$a_1^{(j)} = \frac{\zeta_{11}^{(j)} - \xi_1^2}{\sigma^2}$$

$$a_2^{(j)} = \frac{\sigma^4\zeta_{22}^{(j)} + 2\sigma^2(\xi_1\xi_2 - \xi_3)\zeta_{21}^{(j)} + (\xi_1\xi_2 - \xi_3)^2\zeta_{11}^{(j)} - (\xi_2^2 - \xi_1\xi_3)}{\sigma^4\xi_4 + 2\sigma^2(\xi_1\xi_2 - \xi_3)\xi_3 + (\xi_1\xi_2 - \xi_3)^2\xi_2 - (\xi_2^2 - \xi_1\xi_3)}.$$

In the sections that follow, we will consider the robustness problem (4.1) where the performance measure S is either S_1 or S_3. We assume that under the hypothesis H_i, the true distribution of the observed process is in the class $\mathcal{Q}_i$, which is defined as above by the nominal marginal density $\bar{f}_i$, the parameter ϵ_i, and an r-sequence which has the sum R_i. For given marginal densities, it will be shown that the bivariate distributions defined in (4.5) are in fact least favorable, and this reduces the problem (4.1) to one which involves only the marginal densities. We now give the least favorable marginal

densities, which we will call the *Huber-Strassen least favorable densitites.* These are

$$p_0(x) = \begin{cases} (1-\epsilon_0)\tilde{f}_0(x) & \text{if } \tilde{f}_1(x)/\tilde{f}_0(x) < c'' \\ (1/c'')(1-\epsilon_0)\tilde{f}_1(x) & \text{if } \tilde{f}_1(x)/\tilde{f}_0(x) \geq c'' \end{cases}$$
$$p_1(x) = \begin{cases} (1-\epsilon_1)\tilde{f}_1(x) & \text{if } \tilde{f}_1(x)/\tilde{f}_0(x) > c' \\ c'(1-\epsilon_1)\tilde{f}_0(x) & \text{if } \tilde{f}_1(x)/\tilde{f}_0(x) \leq c' \end{cases} \tag{4.9}$$

where the constants c' and c'' are chosen such that the functions are valid probability densities (i.e. they integrate to 1). The Huber-Strassen densities have appeared frequently as the solution to various minimax robustness problems. Lemma 13 is the basis for many such applications.

Lemma 13. *For $i = 0, 1$, let $\mathcal{P}_i$ be the class of all probability density functions of the form $f = (1-\epsilon_i)\tilde{f}_i + \epsilon_i h$, where $\tilde{f}_i$ is fixed and h is arbitrary, and let Ψ be any convex function. If p_0 and p_1 are the Huber-Strassen least favorable densities corresponding to $\tilde{f}_0$ and $\tilde{f}_1$, then the inequality*

$$\int \Psi\left[\frac{p_1(x)}{p_0(x)}\right] p_0(x)dx \leq \int \Psi\left[\frac{f_1(x)}{f_0(x)}\right] f_0(x)dx$$

holds for all marginal densities $f_0 \in \mathcal{P}_0$ and $f_1 \in \mathcal{P}_1$.

Proof. It has been shown in [7] that the least favorable densities in terms of risk for the classes $\mathcal{P}_0$ and $\mathcal{P}_1$ are the Huber-Strassen densities. The proof then follows as a corollary to Lemma 1 in [15]. □

In addition to the ϵ-contamination classes, the Huber-Strassen densities are also least favorable in terms of risk for at least three other uncertainty classes: the total variation classes [7], bounded classes [17], and p-point classes [18]. Thus Lemma 13 holds as well if the classes $\mathcal{P}_0$, $\mathcal{P}_1$ are both of one of these other three classes.

The main result of this chapter is stated in the following theorem.

Theorem 14. *The least favorable process distributions F_0^*, F_1^* in the classes $\mathcal{Q}_0$, $\mathcal{Q}_1$ are such that their marginal densities are the Huber-Strassen densities (4.9) and their bivariate joint distributions are defined by (4.5).*

4.2 Robustness for S_1

In the first section the idea of minimax robustness was discussed and the problem (4.1) posed without reference to a particular performance measure. We are now ready to find the solution to (4.1) with the performance measure S taken to be S_1 as defined in (2.6). Our first task is to show that for arbitrary but fixed marginal densities, the bivariate disributions defined by (4.5) are in fact least favorable. For $i = 0, 1$, assume that the marginal density f_i is fixed and denote by $\mathcal{R}_i$ the subset of $\mathcal{Q}_i$ containing all the distributions F_i which agree with the fixed marginal density. Let F_i^* denote any such distribution in $\mathcal{R}_i$ having bivariate distributions defined by (4.5). To show that the distributions F_0^* and F_1^* are least favorable, we must show the inequalities (4.1). From the result (4.6) and the fact that S_1 is invariant under the scaling of g, it is clear that the left inequality is an equality for any allowable nonlinearity g. From (4.4) it is clear that the inequality

$$\sigma_1^2(g) \leq (1 + 2R_1)\mathrm{Var}_1\, g(X_1) \tag{4.10}$$

always holds for any distributions in the uncertainty class $\mathcal{Q}_1$. But (4.6) implies that under the distributions F_0^* and F_1^*, equality is obtained in (4.10) for arbitrary g, and in particular equality holds for g^*. These facts imply that the right inequality in (4.1) holds. Thus F_0^* and F_1^* are the least favorable distributions in the classes $\mathcal{R}_0$ and $\mathcal{R}_1$.

Now suppose we define a new performance measure $\hat{S}_1$ by

$$\hat{S}_1(g, f_0, f_1) = \frac{[\mu(g; f_1) - \mu(g; f_0)]^2}{\sigma^2(g; f_1)} \tag{4.11}$$

where we introduce the new notation

$$\mu(g; f) = \int g(x) f(x) dx$$

$$\sigma^2(g; f) = \int g^2(x) f(x) dx - [\mu(g; f)]^2.$$

Such a performance measure depends only on the marginal densities. Suppose that we find g^*, f_0^*, f_1^* which form a saddle point for $\hat{S}_1$, with f_0^* and f_1^* being allowable marginal densities for the classes $\mathcal{Q}_0$ and $\mathcal{Q}_1$. Thus

$$\hat{S}_1(g, f_0^*, f_1^*) \leq \hat{S}_1(g^*, f_0^*, f_1^*) \leq \hat{S}_1(g^*, f_0, f_1). \tag{4.12}$$

Now if we consider the performance measure S_1, and g^*, F_0^*, F_1^*, where F_i^* is a distribution having marginal density f_i^* and bivariate distributions defined by (4.5), we find that we have a saddle point for the classes $\mathcal{Q}_0$ and $\mathcal{Q}_1$. Indeed, we find that

$$S_1(g, F_0^*, F_1^*) = \frac{\hat{S}_1(g, f_0^*, f_1^*)}{(1 + 2R_1)} \leq \frac{\hat{S}_1(g^*, f_0^*, f_1^*)}{(1 + 2R_1)} = S_1(g^*, F_0^*, F_1^*).$$

Furthermore, we have

$$S_1(g^*, F_0^*, F_1^*) = \frac{\hat{S}_1(g^*, f_0^*, f_1^*)}{(1 + 2R_1)} \leq \frac{\hat{S}_1(g^*, f_0, f_1)}{(1 + 2R_1)} = S_1(g^*, F_0', F_1') \leq S_1(g^*, F_0, F_1)$$

where F_i' denotes the process distribution having f_i as the marginal density and bivariate distributions given by (4.5), and F_i denotes an arbitary process distribution which has the marginal density f_i. Our conclusion now is that we need only solve the problem involving the performance measure $\hat{S}_1$ and the marginal densities; that is, we must find

the least favorable marginal densities f_0^* and f_1^* which together with g^* solve the problem (4.12). Then the least favorable process distributions are such that they agree with the marginal densities f_0^*, f_1^* and have bivariate distributions given by (4.5).

Our method will be as follows: First we find the solution g^*, f_0^*, f_1^* to the minimax problem (4.2), where f_0^* and f_1^* are in the classes $\mathcal{Q}_0'$ and $\mathcal{Q}_1'$ which we define to be the classes of all marginal densities which are derived from the classes $\mathcal{Q}_0$, $\mathcal{Q}_1$. Recall that $\mathcal{Q}_i'$ is an ϵ-contamination class with nominal univariate density $\tilde{f}_i$ and parameter ϵ_i. If the problem (4.1) has a solution, then necessarily it must be g^*, f_0^*, f_1^*. Thus at this point we have likely candidates for the robust nonlinearity and the least favorable densities. The second step is to show that the right inequality in (4.1) is satisfied; that is, we must show that

$$\hat{S}_1(g^*, f_0^*, f_1^*) = \inf_{f_0, f_1} \hat{S}_1(g^*, f_0, f_1). \tag{4.13}$$

For given marginals f_0, f_1, the optimal nonlinearity g is given by the solution of the integral equation (2.19) with $m = 0$, and it is easily verified that a solution is given by $g(x) = -[f_0(x)/f_1(x)]$. By Theorem 3, we have that

$$S_1(g, f_0, f_1) = \int g(x)[f_1(x) - f_0(x)]dx = \int \frac{f_0(x)^2}{f_1(x)}dx - 1 \tag{4.14}$$

Thus for the first step, solving (4.2), we must minimize the rightmost integral in (4.14). Lemma 13 applies with $\Psi(x) = x^{-1}$, which is convex. Thus our candidates for the least favorable marginal densities are the Huber-Strassen densities corresponding to the nominals $\tilde{f}_0$ and $\tilde{f}_1$.

We must now show that the right inequality in (4.1) is satisfied by $g^* = -(f_0^*/f_1^*)$, where f_0^* and f_1^* are the Huber-Strassen densities. A complete proof of this result has been published in [16], and therefore only a sketch of the proof will be given here.

Lemma 15, whose proof is given in [13], will be used here as well as in the next section to show that certain functions are convex.

Lemma 15. *If $v_1 > 0$, $v_2 > 0$, and $0 \leq \alpha \leq 1$ then*

$$\frac{\left[\alpha u_1 + (1-\alpha)u_2\right]^2}{\alpha v_1 + (1-\alpha)v_2} \leq \alpha \frac{u_1^2}{v_1} + (1-\alpha)\frac{u_2^2}{v_2}.$$

By virtue of Lemma 15, the performance measure $\hat{S}_1$ is convex in the densities f_0, f_1 for fixed g. This implies that the function $J(\alpha; f_0, f_1) = \hat{S}_1\left[g^*, (1-\alpha)f_0^* + \alpha f_0, (1-\alpha)f_1^* + \alpha f_1\right)$ is convex in α. Furthermore, a necessary and sufficient condition for (4.13) to hold is that $\frac{d}{d\alpha}J(\alpha; f_0, f_1)\Big|_{\alpha=0} \geq 0$ for arbitrary f_0, f_1. However, it is also true from considering (4.14) that f_0^* and f_1^* minimize the functional $T[f_0, f_1] = \int(f_0^2/f_1)$. Since T is also convex in f_0, f_1, we must have also the condition that $\frac{d}{d\alpha}T\left[(1-\alpha)f_0^* + \alpha f_0, (1-\alpha)f_1^* + \alpha f_1\right]_{\alpha=0} \geq 0$. By considering these two derivatives, it can be shown that the condition that f_0^*, f_1^* minimize T is equivalent to the condition that they minimize $\hat{S}_1(g^*, f_0, f_1)$. A similar proof is given in greater detail in the next section for the performance measure S_3.

4.3 Robustness for S_3

We will obtain in this section essentially the same result for the performance measure S_3 as that obtained in the preceding section for the performance measure S_1. The first task is to show that the problem of finding the least favorable distributions again reduces to a problem involving only the marginal densities. Define a new performance measure

$$\hat{S}_3(g, f_0, f_1) = \frac{\left[\mu(g; f_1) - \mu(g; f_0)\right]^2}{\sigma^2(g; f_0) + A\sigma^2(g; f_1)}, \tag{4.15}$$

where $A = [(1 + 2R_1)/(1 + 2R_0)]$. Such a performance measure depends only on the marginal distributions f_0, f_1. If F_0, F_1 are process distributions which have marginal distributions f_0, f_1 and bivariate distributions defined by (4.5), then we have the relation

$$S_3(g, F_0, F_1) = \frac{\hat{S}_3(g, f_0, f_1)}{(1 + 2R_0)}.$$

A series of equalities and inequalities similar to those at the beginning of Section 4.2 can be used to show that the problem (4.1) for the performance measure S_3 reduces to the problem involving only the performance measure $\hat{S}_3$. Thus if one finds the least favorable marginal distributions f_0^*, f_1^* and the minimax robust nonlinearity g^* for the performance measure $\hat{S}_3$, then the minimax problem for S_3 is solved by taking the least favorable process distributions to be such that the marginal distributions are f_0^*, f_1^* and the bivariate distributions are given by (4.5).

The integral equation which yields the optimal nonlinearity for $\hat{S}_3$ is similar to (3.27) with $m = 0$ except for the coefficient A:

$$g(x) = \frac{f_1(x) - f_0(x)}{f_0(x) + Af_1(x)} + \int \left[\frac{f_0(x)f_0(y) + Af_1(x)f_1(y)}{f_0(x) + Af_1(x)}\right] g(y)dy. \tag{4.16}$$

We have immediately the form of the solution

$$g(x) = \frac{B_0 f_0(x) + B_1 f_1(x)}{f_0(x) + Af_1(x)} \tag{4.17}$$

where $B_0 = \mu_0 - 1$ and $B_1 = A\mu_1 + 1$. If we consider the linear system of equations

$$\mu_0 = B_0 + 1 = \int g(x)f_0(x)dx$$

$$\mu_1 = \frac{1}{A}(B_1 - 1) = \int g(x)f_1(x)dx$$

with B_0, B_1 as the unknowns, then we find that the system is singular, and consequently we may assign to B_0 the arbtrary value 0. This implies that

$$B_1 = \left[\int \frac{f_0(x)f_1(x)}{f_0(x) + Af_1(x)}dx\right]^{-1}. \tag{4.18}$$

If g is the optimal nonlinearity which is matched to f_0, f_1, then we know that

$$\begin{aligned}\hat{S}_3(g, f_0, f_1) &= \int g(x)[f_1(x) - f_0(x)]dx \\ &= B_1(f_0, f_1) \int \frac{f_1(x)}{f_0(x) + Af_1(x)}[f_1(x) - f_0(x)]dx.\end{aligned} \tag{4.19}$$

where we have written B_1 as a function of f_0 and f_1 to remind us of the relation (4.18). Lemma 13 applies to the integral in (4.19) with $\Psi(x) = x(x-1)/(x+1)$, which is convex, so that the integral in (4.19) is minimized by the Huber-Strassen densities. Lemma 13 also applies to the integral in (4.18). In this case $\Psi(x) = x/(Ax+1)$, which is concave, so that by applying the lemma to the negative of the integral (since $-\Psi$ is convex) we find that this integral is maximized by the Huber-Strassen densities. $B_1(f_0, f_1)$ therefore is minimized. Thus our candidates for the least favorable marginal densities are the Huber-Strassen densities. The right inequality in (4.1) will now be proved.

The following inequalities, which depend on the fact that $\sigma^2(g; f)$ is concave in f and on Lemma 15, demonstrate that $\hat{S}_3(g, f_0, f_1)$ is convex in f_0 and f_1 for fixed g. With $\beta = (1 - \alpha)$ we have

$$\begin{aligned}\hat{S}_3(g, \beta\bar{f}_0 + \alpha f_0, \beta\bar{f}_1 + \alpha f_0) &= \frac{[\mu(g; \beta\bar{f}_1 + \alpha f_1) - \mu(g; \beta\bar{f}_0 + \alpha f_0)]}{\sigma^2(g; \beta\bar{f}_0 + \alpha f_0) + A\sigma^2(g; \beta\bar{f}_1 + \alpha f_1)} \\ &\leq \frac{\left[\beta\{\mu(g; \bar{f}_1) - \mu(g; \bar{f}_0)\} + \alpha\{\mu(g; f_1) - \mu(g; f_0)\}\right]^2}{\beta[\sigma^2(g; \bar{f}_0) + A\sigma^2(g; \bar{f}_1)] + \alpha[\sigma^2(g; f_0) + A\sigma^2(g; f_1)]} \\ &\leq \beta\hat{S}_3(g, \bar{f}_0, \bar{f}_1) + \alpha\hat{S}_3(g, f_0, f_1)\end{aligned}$$

Define the function

$$J(\alpha; f_0, f_1) = \hat{S}_3[g^*, (1-\alpha)f_0^* + \alpha f_0, (1-\alpha)f_1^* + \alpha f_1] \qquad 0 \leq \alpha \leq 1$$

where f_0^* and f_1^* are the Huber-Strassen least favorable densities and g^* is the optimal nonlinearity matched to f_0^*, f_1^*. Certainly J is convex in α if $\hat{S}_3$ is convex in f_0 and f_1.

Now the right inequality in (4.1) holds if and only if

$$J(\alpha; f_0, f_1) \geq J(0; f_0, f_1) \tag{4.20}$$

for all α in the interval $[0,1]$, and since J is convex in α, (4.20) holds if and only if we have the condition

$$\frac{d}{d\alpha} J(\alpha; f_0, f_1)\Big|_{\alpha=0} \geq 0. \tag{4.21}$$

If we take the derivative of $J(\alpha; f_0, f_1)$ and set $\alpha = 0$. Then we have

$$\begin{aligned}\frac{d}{d\alpha} J(\alpha; f_0, f_1)\Big|_{\alpha=0} &= 2\int g^*(f_1 - f_0 - f_1^* + f_0^*) - \int (g^*)^2(f_0 + Af_1) + \int (g^*)^2(f_0^* + Af_1^*) \\ &\quad + 2\Big[\int g^* f_0^*\Big]\Big[\int g^*(f_0 - f_0^*)\Big] + 2A\Big[\int g^* f_1^*\Big]\Big[\int g^*(f_1 - f_1^*)\Big] \\ &= 2B_1\Big[\int g^* f_1 - \int g^* f_1^*\Big] + \int (g^*)^2(f_0^* + Af_1^*) - \int (g^*)^2(f_0 + Af_1).\end{aligned} \tag{4.22}$$

We can now show that (4.21) holds by considering the function

$$T[f_0, f_1] = \int \frac{f_1(x)^2}{f_0(x) + Af_1(x)} dx \tag{4.23}$$

which by Lemma 13 is minimized by the Huber-Strassen densities. Define

$$K(\alpha; f_0, f_1) = T\big[(1-\alpha)f_0^* + \alpha f_0, (1-\alpha)f_1^* + \alpha f_1\big].$$

It follows from Lemma 15 that T is convex in f_0, f_1. By the same reasoning as before, then, we conclude that f_0^*, f_1^* minimize T if and only if

$$\frac{d}{d\alpha} K(\alpha; f_0, f_1)\Big|_{\alpha=0} \geq 0. \tag{4.24}$$

The final step in our proof is to show that the inequality (4.24) implies the inequality (4.21). Define $p_\alpha^{(i)} = (1-\alpha)f_i^* + \alpha f_i$ for $i = 0, 1$ and $0 \le \alpha \le 1$, so that we have

$$K(\alpha; f_0, f_1) = \int \frac{(p_\alpha^{(1)})^2}{p_\alpha^{(0)} + A p_\alpha^{(1)}}, \tag{4.25}$$

and thus

$$\frac{d}{d\alpha} K(\alpha; f_0, f_1)\Big|_{\alpha=0} = \lim_{\alpha \to 0} \int \frac{1}{\alpha} \left[\frac{(p_\alpha^{(1)})^2}{p_\alpha^{(0)} + A p_\alpha^{(1)}} - \frac{(p_0^{(1)})^2}{p_0^{(0)} + A p_0^{(0)}} \right].$$

The derivative of the integrand in (4.25) is

$$\frac{d}{d\alpha} \left[\frac{(p_\alpha^1)^2}{p_\alpha^0 + A p_\alpha^1} \right]_{\alpha=0} = 2\frac{f_1^*}{f_0^* + A f_1^*}(f_1 - f_1^*) + \left(\frac{f_1^*}{f_0^* + A f_1^*} \right)^2 [(f_0^* + A f_1^*) - (f_0 + A f_1)].$$

To differentiate K we must justify the interchanging of the integration and differentiation operations. The convexity of K as a function of α implies the inequalities

$$\begin{aligned} 2\frac{f_1^*}{f_0^* + A f_1^*}(f_1 - f_1^*) + \left(\frac{f_1^*}{f_0^* + A f_1^*} \right)^2 & [(f_0^* + A f_1^*) - (f_0 + A f_1)] \\ & \le \frac{1}{\alpha} \left[\frac{(p_\alpha^{(1)})^2}{p_\alpha^{(0)} + A p_\alpha^{(1)}} - \frac{(p_0^{(1)})^2}{p_0^{(0)} + A p_0^{(1)}} \right] \\ & \le \frac{(p_1^{(1)})^2}{p_1^{(0)} + A p_1^{(1)}} - \frac{(p_0^{(1)})^2}{p_0^{(0)} + A p_0^{(1)}} \end{aligned} \tag{4.26}$$

The right quantity in (4.26) is integrable, and the middle quantity converges pointwise monotonically to the left quantity as $\alpha \to 0$ because of the convexity of K. The monotone convergence theorem then permits the interchange of the differentiation and the integration, and we have

$$\begin{aligned} & \frac{d}{d\alpha} K(\alpha; f_0, f_1)\Big|_{\alpha=0} = \\ & \int \left\{ 2\frac{f_1^*}{f_0^* + A f_1^*}(f_1 - f_1^*) + \left(\frac{f_1^*}{f_0^* + A f_1^*} \right)^2 [(f_0^* + A f_1^*) - (f_0 + A f_1)] \right\}. \end{aligned} \tag{4.27}$$

Now if we compare equations (4.22) and (4.27), then we see that (compare (4.17))

$$\frac{d}{d\alpha} J(\alpha; f_0, f_1)\Big|_{\alpha=0} = \frac{d}{d\alpha} B_1^2 K(\alpha; f_0, f_1)\Big|_{\alpha=0}$$

and thus conditions (4.21) and (4.24) are equivalent.

CHAPTER 5

NUMERICAL RESULTS AND CONCLUSION

5.1 Description of the Examples

In the work of the preceding chapters, we attempted to justify the use of the various performance measures by showing that a large value of a performance measure results good performance as determined by the actual error probabilities. Such an approach is necessary since the performance measures are mathematically tractable, whereas the error probabilities themselves are not. The error probabilities, however, can be estimated by simulation on a digital computer. Such simulation results, presented in this chapter for several different examples, will complete this work.

There are two questions concerning which we might like to gain some insight as a result of these computer simulations. First, and perhaps foremost, is the question about the validity of the assumptions which were made in justifying the various performance measures. In particular, we assumed that the distribution of the test statistic was approximately Gaussian, and in fact, under the hypotheses of Theorem 1 or Theorem 12 the distribution of the normalized test statistic converges to a Gaussian distribution as the sample size approaches infinity. Our tests shall have finite sample sizes, however,

and therefore the effect of the finite sample size should be examined. The second question which warrants our attention is of a philosophical nature. The processes between which we wish to discriminate are dependent, and thus they necessarily involve memory. However, the tests with which we wish to perform such discrimination are memoryless, and therefore it is not clear to the intuition that such a scheme can work effectively. In the case of mixing processes, where there is "asymptotic independence," we know that memoryless discrimination is possible, and that the performance will improve as the sample size increases. Our concern here should be the improvement of a given memoryless discriminator over the discriminator which is designed under the assumption that the processes are iid (i.e. the LRT (1.2)).

In all of the examples which are presented here, the marginal densities will be the same throughout, and the varying parameters will be the time constants of the dependency lengths and the sample sizes of the tests. We assume that the process has a Rayleigh distribution with parameter $\theta = 4$ under hypothesis H_0 and a lognormal distribution with parameters $\lambda_1 = 0.8, \lambda_2 = 0.25$ under hypothesis H_1. The Rayleigh and lognormal densities are given in Appendices A and B, respectively. These distributions are have found application in radar discrimination problems, and indeed such has been the motivation behind this research. The n-dimensional Rayleigh density is such that it generally lacks a closed form expression when $n > 2$, and thus an LRT is not feasible. The parameters ρ_j which appear in the expressions for both the Rayleigh bivariate densities (A2) and the lognormal bivariate densities (B4) are actually the correlation coefficients of the underlying Gaussian process(es). In each of the examples of this chapter, we shall assume that the values of the ρ_j parameters are given by exponentially decaying sequences which are determined by a time constant τ_i. Thus under hypothesis

H_i, we have $\rho_j = \exp(-j/\tau_i)$, where ρ_j is the parameter in the density f_i^j. The time constants will be varied in the different examples to reveal the effects of varying degrees of dependency on the various test statistics.

Several comments concerning the choices for the aforementioned parameters are in order. First, the parameters for the marginal densities were chosen to match as closely as possible the two densities involved. More precisely, the parameters are such that $\mathrm{E}_0 X_1 = \mathrm{E}_1 X_1$, and $\mathrm{E}_0 X_1^2 = \mathrm{E}_1 X_1^2$; that is, the first and second moments agree. The graphs of the two marginal densities can be observed in Figures 1 and 2. While observing the linear plots in Figure 1, it seems that this is a relatively difficult discrimination problem; however, logarithmic plots in Figure 2 reveal that there is a great deal of discrimination capability in the tail regions, the Rayleigh density f_0 having a much heavier tail to the left and the lognormal density f_1 having a heavier tail to the right. Second, by taking the sequences of ρ-parameters to be exponential sequences, the underlying Gaussian processes become Markov processes, and thus methods for generating the processes on a computer become relatively simple.

As mentioned above, the parameters for the marginal densities shall remain the same for each of the specific examples considered. The time constants, however, will be varied in the different examples. We shall assign a label E_i to each of the examples for easy reference. Table 1 lists the parameters for each of the examples.

5.2 The Calculation of the Nonlinearities

For each of the examples $E_1, \ldots, E_5$, we shall compare the performance of five different nonlinearities $g_i, i = 0, \ldots, 4$. We denote by g_i, for $0 \leq i \leq 3$, the optimal

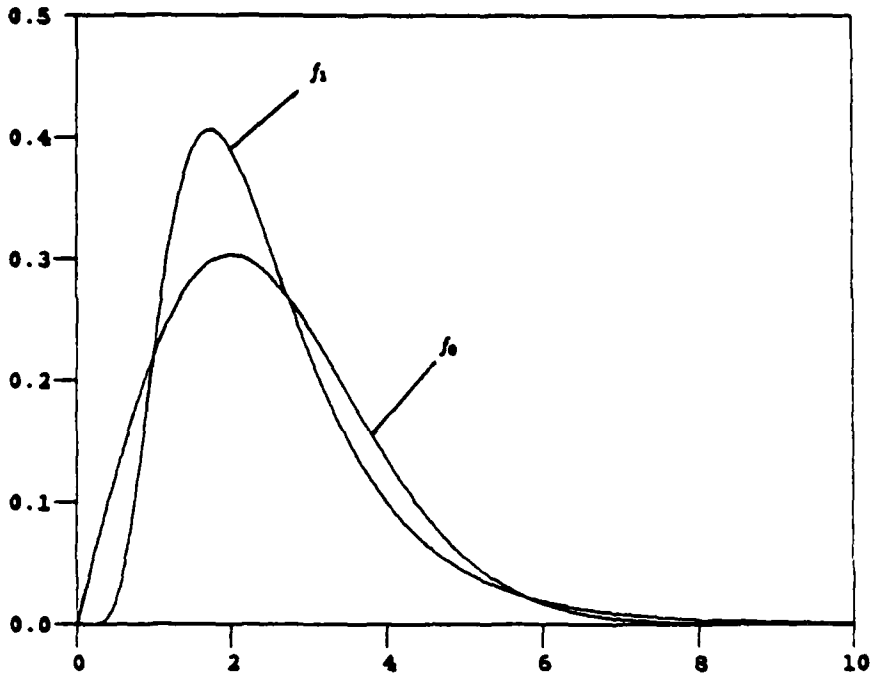

Figure 1. Linear graphs of the marginal densities f_0, f_1.

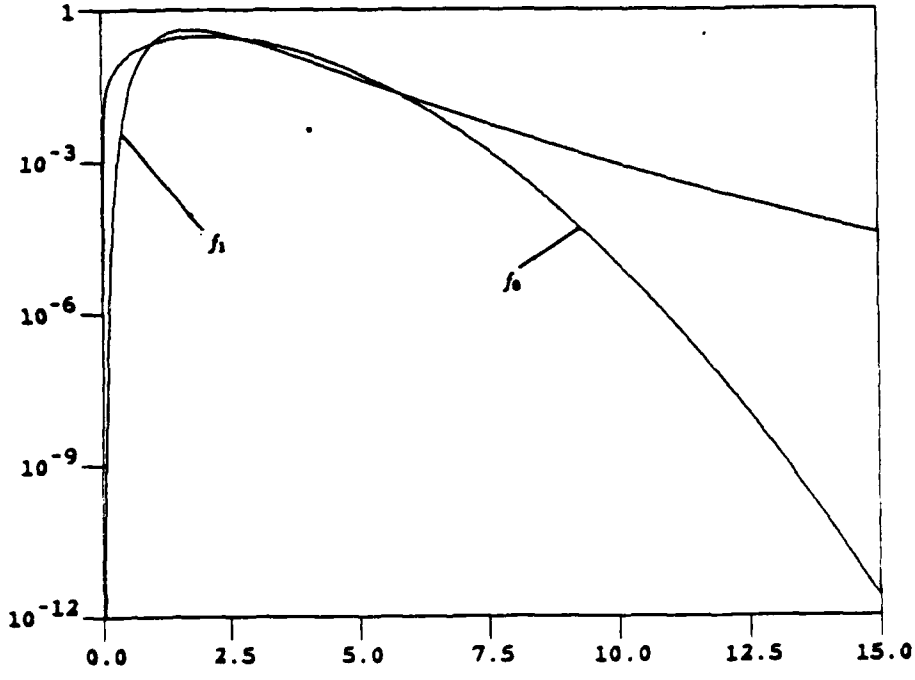

Figure 2. Logarithmic graphs of the marginal densities f_0, f_1.

Table 1. Parameters for the examples.

Example	τ_0	τ_1	m_0	m_1	n
E_1	13.0288	13.0288	60	60	1000
E_2	13.0288	130.288	60	600	1000
E_3	130.288	13.0288	600	60	1000
E_4	130.288	130.288	600	600	1000
E_5	13.0288	130.288	60	600	100

nonlinearity for the performance measure S_i, which is consistent with our usage in the preceding chapters. We also denote by g_4 the optimal iid nonlinearity given by $g_4 = \log(f_1/f_0)$ (cf. (1.2)). Thus we have also five different test statistics $T_i = \sum_{k=1}^{n} g_i(x_k)$, $i = 0, \ldots, 4$. The nonlinearity g_4 is computed easily since it has a closed form solution. To obtain the others, the corresponding integral equations from Chapters 2 and 3 must be solved.

Several issues must be considered in the numerical calculation of the nonlinearities. First, because the integral equations are derived for m-dependent processes, we must assign a value to m. Our criterion for doing so is to select m so that $\rho_m \leq \rho_{\min}$. Thus we have two values m_0, m_1 corresponding to the processes under the two hypotheses H_0, H_1. In the results presented here, we have $\rho_{\min} = 0.01$. These results were tested by decreasing the value of $\rho_{\min}$, or equivalently, increasing the value of m_i and it was found that the numerical results were unchanged, thus corroborating Theorems 6 and 11. The second issue is the choice of a finite interval $[x_{\min}, x_{\max}]$ over which the integration is to be performed. This amounts to truncating the densities, and is of a special concern for

the nonlinearities g_0 and g_1 since $f_0(x)/f_1(x)$ is unbounded as $x \to \infty$ and $f_1(x)/f_0(x)$ is unbounded as $x \to 0$. Thus for these cases the absolute term in the integral equation does not have a finite second moment, and it is therefore manditory that the tails of the densities be modified or truncated for the problem to be well defined as a Fredholm equation. In other words, the condition (a) at the beginning of Section 2.3 is not satisfied unless the tails of the densities are modified by truncation or some other method. For the problems here, the interval $[x_{\min}, x_{\max}]$ was chosen so that under either hypothesis $P\{X_1 < x_{\min}\} < \epsilon$ and $P\{X_1 > x_{\max}\} < \epsilon$, where $\epsilon = 5 \times 10^{-5}$. This resulted in $x_{\min} = 0.02$ and $x_{\max} = 15.7$.

The most direct method for solving a Fredholm equation is to approximate the integral with a numerical quadrature formula, and thereby transform the problem into a system of linear equations. In the method used here the quadrature formula was a composite Simpson's rule with $N = 301$ nodes. The argument goes as follows. First approximate the integral by the weighted sum thus:

$$g(x) \approx \frac{f_1(x) - f_0(x)}{f_1(x)} + \sum_{i=1}^{N} K(x, x_i) g(x_i) w_i. \tag{5.1}$$

We find our numerical solution by solving the N linear equations

$$u_j = \frac{f_1(x_j) - f_0(x_j)}{f_1(x_j)} + \sum_{i=1}^{N} K(x_j, x_i) u_i w_i \qquad j = 1, \ldots, N \tag{5.2}$$

for the N unknowns $u_i, i = 1, \ldots, N$. If the numerical integration in (5.1) is reasonably accurate for all the x values in the interval $[x_{\min}, x_{\max}]$, then $g(x_i)$ will solve a linear system of equations similar to those in (5.2) but with slightly perturbed coefficients. Therefore, provided the coefficient matrix is not ill-conditioned, the solution of (5.2) will give a reasonably approximate solution to the integral equation. In fact, Fredholm

actually proved that the such approximations converge to the solution as N approaches infinity. Obtaining a numerical solution for the integral equation (3.15) requires a little more effort, in that one must solve the linear equations for several different values of τ until the sequence of τ values appears to be near its limit. In the problems solved here, the initial value was $\tau = 1$ and convergence to within 10^{-4} occurred after 10–23 iterations for the four examples. For any of the integral equations, if the values of m_0 and m_1 are large, then typically the vast portion of CPU time is spent in initializing the matrix for the linear system because of the sums in the kernels. For the integral equation (3.15), where the linear system must be solved several times, it is economical to save the values of the sums.

Shown in Figures 3–6 are the graphs of the numerically computed nonlinearities for the problem E_2. Figure 3 displays the nonlinearities g_0 and g_4, while Figure 4, which is drawn to a much smaller scale than Figure 3, displays the nonlinearities g_1, g_2, and g_3. Figures 5 and 6 are semilogarithmic plots which show the right and left tails of the nonlinearities. The tail behavior is of concern because, as we noted by observing Figure 2, this is where most of the discrimination capability lies. We might try to predict the performance of each of the test statistics by observing the shapes of the corresponding nonlinearities. For g_0, the heavy tail to the right will cause a separation of the means $\mu_0(g_0)$ and $\mu_1(g_0)$ at the expense of making $\sigma_1^2(g_0)$ rather large. Of course, $\sigma_0^2(g_0)$ will not be effected in a serious way by the right tail because, as can be seen in Figure 2, f_0 places little mass in that region. We notice the reverse situation for g_1, where the heavy tail on the left should create a separation of $\mu_0(g_1)$ and $\mu_1(g_1)$ at the expense of making $\sigma_0^2(g_1)$ large. Because g_4 has heavy tails on both the left and the right, we would expect $\mu_1(g_4) - \mu_0(g_4)$ to be large, as well as $\sigma_0^2(g_4)$ and $\sigma_1^2(g_4)$. Finally, we note that g_2 and g_3

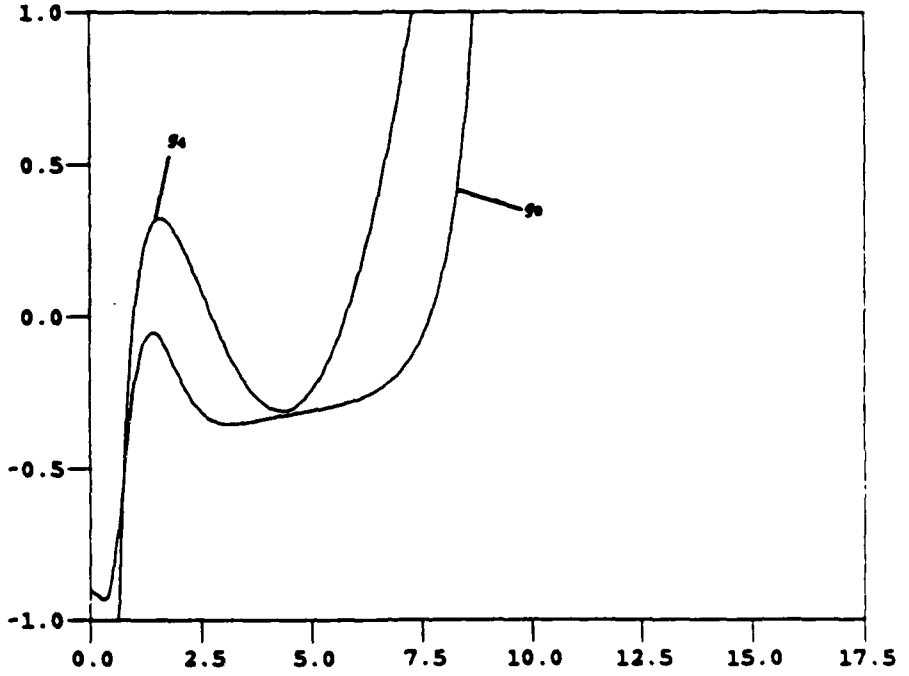

Figure 3. Linear graphs of the nonlinearities g_0 and g_1 for Example E_2.

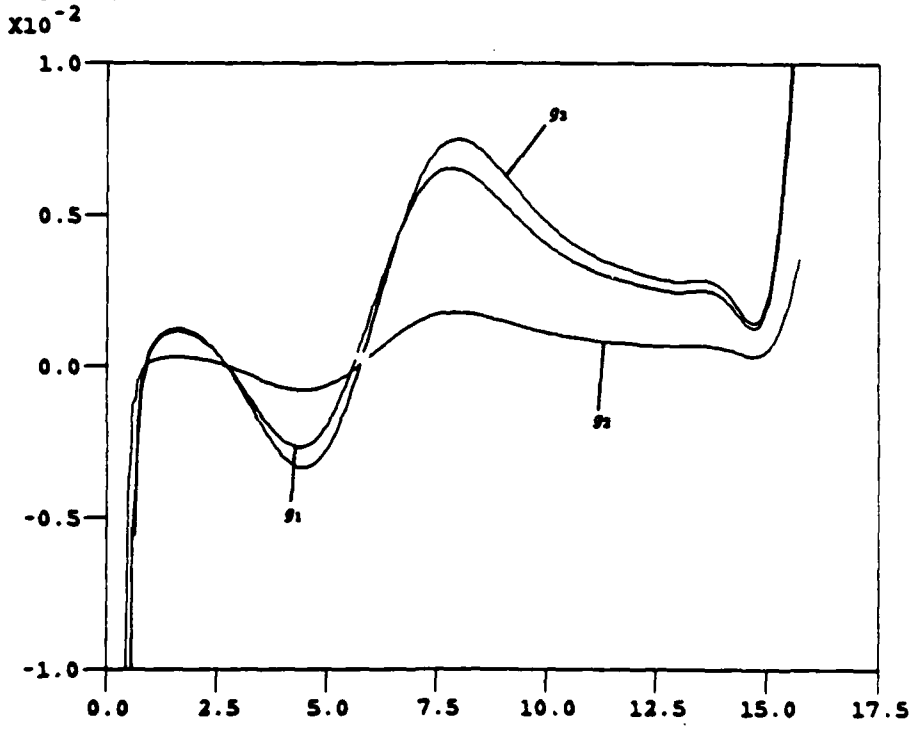

Figure 4. Linear graphs of the nonlinearities g_1, g_2, and g_3 for Example E_2.

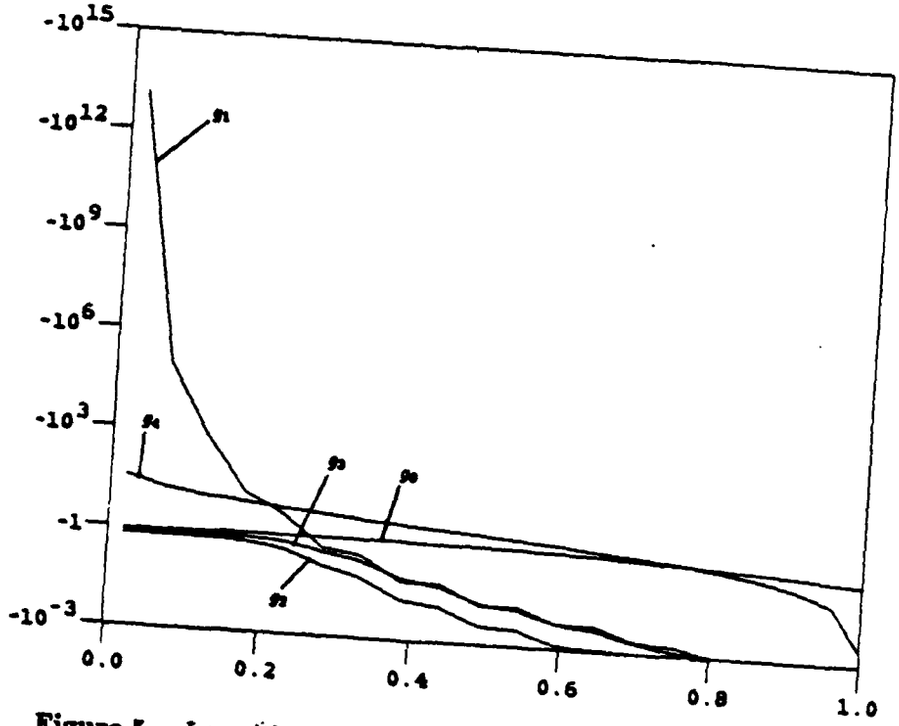

Figure 5. Logarithmic graphs of the left tails of the nonlinearities g_i, $i = 0, \ldots, 4$ for Example E_2.

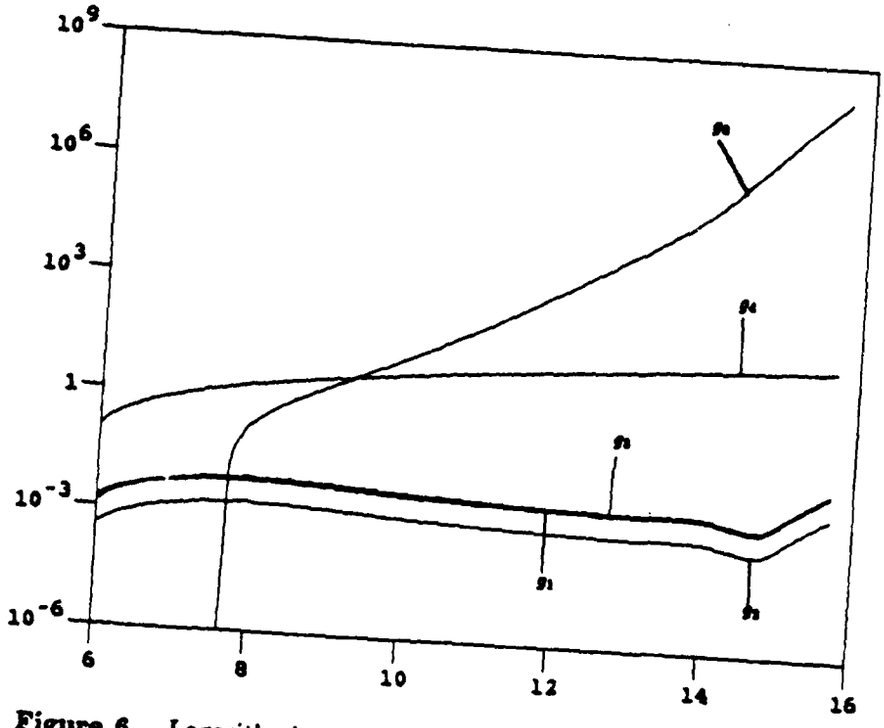

Figure 6. Logarithmic graphs of the right tails of the nonlinearities g_i, $i = 0, \ldots, 4$ for Example E_2.

Table 2. Values of μ_i and σ_i evaluated at the each of the nonlinearities for Example E_2.

	μ_0	σ_0	μ_1	σ_1
g_0	−2.9643e−01	3.1015e+01	9.6166e+02	9.3037e+05
g_1	−1.2197e+09	1.3087e+11	9.1490e−06	3.4924e+04
g_2	−4.9429e−03	5.3269e−02	5.6676e−07	1.7041e−02
g_3	−8.7036e−03	7.9951e−02	1.7709e−06	4.8095e−02
g_4	−1.6660e−01	1.4059e+00	8.0028e−02	5.7433e+00

do not have heavy tails to either the left or right, and thus we expect $\mu_1 - \mu_0$ to be small as well as σ_0^2 and σ_1^2 for each case. Table 2, which lists the values of these moments for each of the nonlinearities for example E_2, shows that such predictions are accurate. One final comment concerning the shapes of the nonlinearities is worth mentioning. Because of the lopsided nature of the nonlinearities g_0 and g_1, the distributions of $g_0(X_1)$ and $g_1(X_1)$ will be skewed to the right and left, respectively. Thus convergence of the sums T_0 and T_1 to a Gaussian distribution will be slow. On the other hand, because the nonlinearities g_2 and g_3 are relatively small in magnitude and "balanced," the convergence of T_2 and T_3 to a Gaussian should be more rapid. The heavy tails on both the left and right of g_4 should cause $g_4(X_1)$ to be skewed to the left under H_0 and skewed to the right under H_1. Thus convergence of T_4 to a Gaussian distribution should also be rather slow. The same general phenomena occur for the other examples as well.

In Tables 3–6 are listed the values of the performance measures evaluated at each of the nonlinearities. We observe for each case that the numerical solutions are consistent with the goal that g_i maximize S_i for $i = 0, 1, 2, 3$.

Table 3. Values of the performance measures evaluated at each of the nonlinearities for Example E_1.

	S_0	S_1	S_2	S_3
g_0	9.6196e+02	1.0008e−05	1.0006e−05	1.0008e−05
g_1	8.6869e−05	7.0651e+10	8.6869e−05	8.6869e−05
g_2	2.2580e−02	9.3601e−02	1.0155e−02	1.8191e−02
g_3	3.0055e−02	5.4255e−02	9.8783e−03	1.9341e−02
g_3	3.0775e−02	2.4026e−02	6.7720e−03	1.3493e−02

Table 4. Values of the performance measures evaluated at each of the nonlinearities for Examples E_2 and E_5.

	S_0	S_1	S_2	S_3
g_0	9.6196e+02	1.0690e−06	1.0690e−06	1.0690e−06
g_1	8.6869e−05	1.2197e+09	8.6869e−05	8.6869e−05
g_2	8.6121e−03	8.4155e−02	4.9434e−03	7.8126e−03
g_3	1.1856e−02	3.2762e−02	4.6221e−03	8.7054e−03
g_4	3.0775e−02	1.8440e−03	1.1901e−03	1.7397e−03

Table 5. Values of the performance measures evaluated at each of the nonlinearities for Example E_3.

	S_0	S_1	S_2	S_3
g_0	1.7417e+02	4.9968e−06	4.9951e−06	4.9968e−06
g_1	8.2661e−05	7.0651e+10	8.2661e−05	8.2661e−03
g_2	5.5079e−03	2.8253e−02	2.6506e−03	4.6093e−03
g_3	6.5927e−03	1.8214e−02	2.5700e−03	4.8406e−03
g_3	4.4120e−03	2.4026e−02	2.1620e−03	3.7275e−03

Table 6. Values of the performance measures evaluated at each of the nonlinearities for Example E_4.

	S_0	S_1	S_2	S_3
g_0	1.7417e+02	8.9601e−07	8.9588e−07	8.9601e−07
g_1	8.2661e−05	1.2197e+9	8.2661e−05	8.2661e−05
g_2	1.8637e−03	6.0845e−02	1.3499e−03	1.8083e−03
g_3	3.0956e−03	8.4433e−03	1.2010e−03	2.2651e−03
g_3	4.4120e−03	1.8440e−03	6.8021e−04	1.3005e−03

5.3 Simulation Results

Figures 7–11 contain the graphs of the receiver operating characteristic curves for each of the examples. These were generated by tabulating the results of 10,000 simulations under each hypothesis. Since the values of the nonlinearities were computed for only those values in the interval $[x_{\min}, x_{\max}]$, a method had to be chosen to deal with those observations outside the interval, and this was resolved by limiting the observations, so that a value outside the interval was reset to $x_{\min}$ or $x_{\max}$, whichever was the closest. The reasoning behind such a method is that real world observations are in fact limited since the instruments which make the measurements are limited. The endpoints $x_{\min}$ and $x_{\max}$ were selected so that the probability of an observation being larger than $x_{\max}$, for example, would be less than 5×10^{-5}, and the actual proportion of observations which occured beyond $x_{\max}$ or below $x_{\min}$ in the simulations proved to be consistent with this probability. Thus the effect of this limiting is practically negligible.

The ROCs in Figures 7–11 are plots of the error probability P_1 versus the error probability P_0 on logarithmic scales. Our main concern shall be the minimax point of the ROC, or that point where $P_0 = P_1$. This region occurs along the diagonal which extends from the lower left corner to the upper right corner of the graph and gives us an ordering of the nonlinearities. The approximate values of P_1 (and hence also P_0) at the minimax point are listed in Table 7. From Figure 7, which corresponds to the example E_1, we see that the iid nonlinearity g_4 performs uniformly better than the others, which is to be expected because the dependency under either hypothesis is relatively weak. The ordering, from best to worst, continues with g_3, g_2, g_1, and finally g_0.

Figure 8 corresponds to E_2, were there is a relatively strong dependency under

Table 7. Approximate values of P_0 (and P_1) at the minimax region of the ROCs for each of the nonlinearities as estimated through computer simulation.

Example	g_0	g_1	g_2	g_3	g_4
E_1	4.1e−03	2.2e−03	1.1e−03	9.0e−04	7.0e−04
E_2	1.1e−01	2.6e−03	4.3e−03	6.5e−03	3.5e−02
E_5	4.0e−02	2.8e−02	2.4e−02	2.8e−02	2.3e−02
E_4	1.8e−01	4.9e−02	4.4e−02	6.9e−02	9.8e−02
E_3	3.7e−01	1.6e−01	1.7e−01	2.4e−01	1.7e−01

the hypothesis H_1. The ordering is g_1, g_2, g_3, g_4, and g_0, a result which is rather pleasing to the intuition. Since there is memory in the observations, the nonlinearity g_4 which is designed under a no memory assumption performs relatively poorly. The nonlinearity g_1, however, was designed to minimize σ_1^2, which essentially captures all of the dependency under H_1. Thus g_1 achieves its relatively good performance by minimizing the effects of the dependency, and this may perhaps be the only way to handle dependency when a memoryless discriminator is to be used.

With this concept in mind, we now examine Figure 9, corresponding to E_3 in which the dependency under H_0 is relatively strong. Although we would expect a reverse of the situation of E_2, we find again that the ordering at the minimax point is g_4, g_2, g_3, g_1, g_0, which is like that of E_1 except that the positions of g_2 and g_3 are reversed. There is not a large difference in the performance from best to worst, and in fact g_1 and g_3 are actually tied for the third position. As we proceed to the left of the curves from the minimax region, we find that there is a region where g_1 performs best, and finally a region where g_0 performs best. If we examine the situation a little more carefully, we may

also have an intuitively pleasing explanation for this result. The correlation with which we are actually dealing is that of the underlying Gaussian process(es). The ρ-parameters in the bivariate densities in the appendices are the actual correlation coefficients of the Gaussian processes, but are related to the correlation coefficients of the Rayleigh and lognormal processes in a one-to-one manner. Denote by $\rho_R(\rho)$ the correlation coefficient of the Rayleigh density as a function of the parameter ρ from the bivariate density. Let $\rho_\ell(\rho)$ denote this function for the lognormal density. Then ρ_ℓ has the explicit form

$$\rho_\ell(\rho) = \frac{e^{\rho\lambda_2} - 1}{e^{\lambda_2} - 1}$$

and we note that derivative of ρ_ℓ at 0 is positive. In fact, for $\lambda_2 = 0.25$ the derivative at 0 has the approximate value 0.88. Although there is no closed form expression for ρ_R, one can show that the derivative at 0 is 0. The implication from this is that the correlation of the Rayleigh density for a given value of the parameter ρ is much less than that of the lognormal density. This makes intuitive sense since the Rayleigh process involves the sum of two Gaussian processes, whereas the lognormal process involves only one. Thus for E_3, there is an increase in the dependency under H_0 compared to that for E_1, but this increase is perhaps not so significant that g_0 would perform better than g_4, as we might expect. What we observe, however, is a degradation of the results for E_1 due to the increase in the dependency under H_0.

We have the ROCs corresponding to E_4 in Figure 10, where a nearly uniform ordering is g_1, g_2, g_3, g_4, g_0. This ordering is precisely that of E_2, although there is not as large a difference in performance here. This is the situation in which the dependency under both hypotheses has been increased from E_1. From the discussion of the preceding paragraph, we know that the dependency under H_1 is stronger than that under H_0. Thus

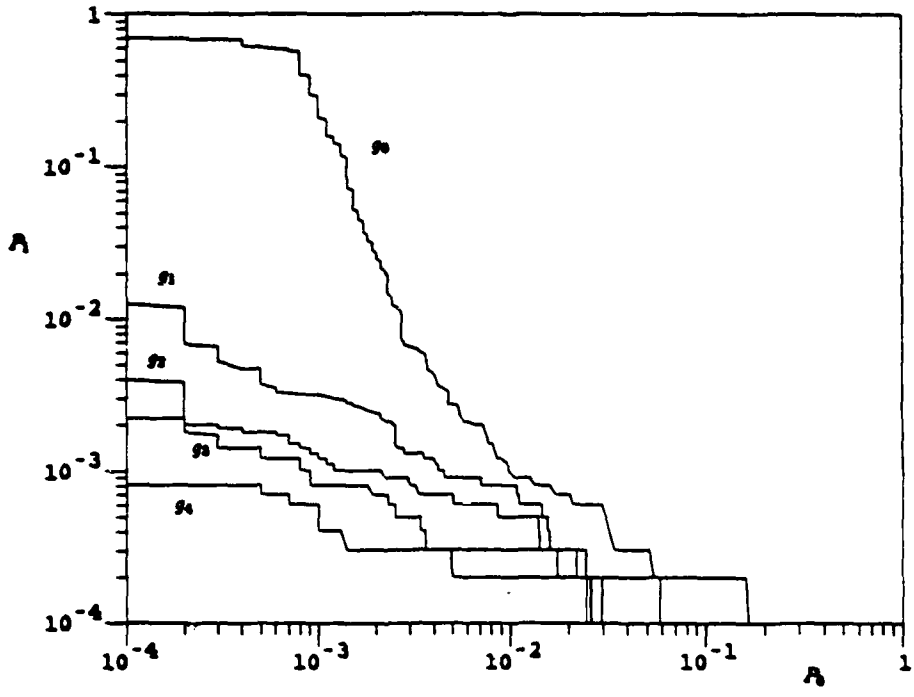

Figure 7. Receiver operating characteristics (ROCs) for Example E_1.

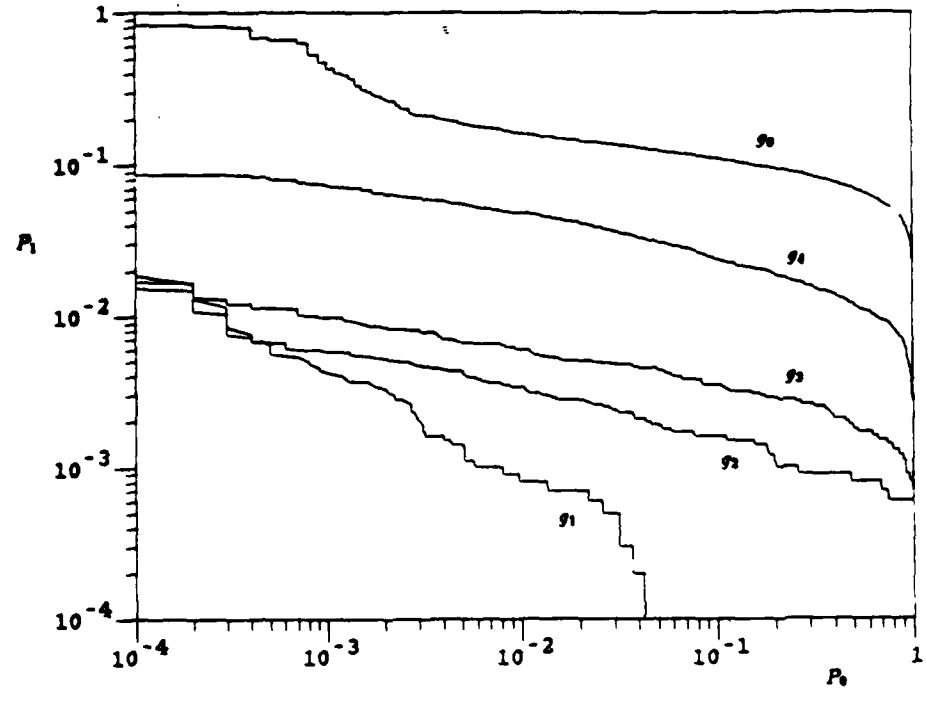

Figure 8. ROCs for Example E_2.

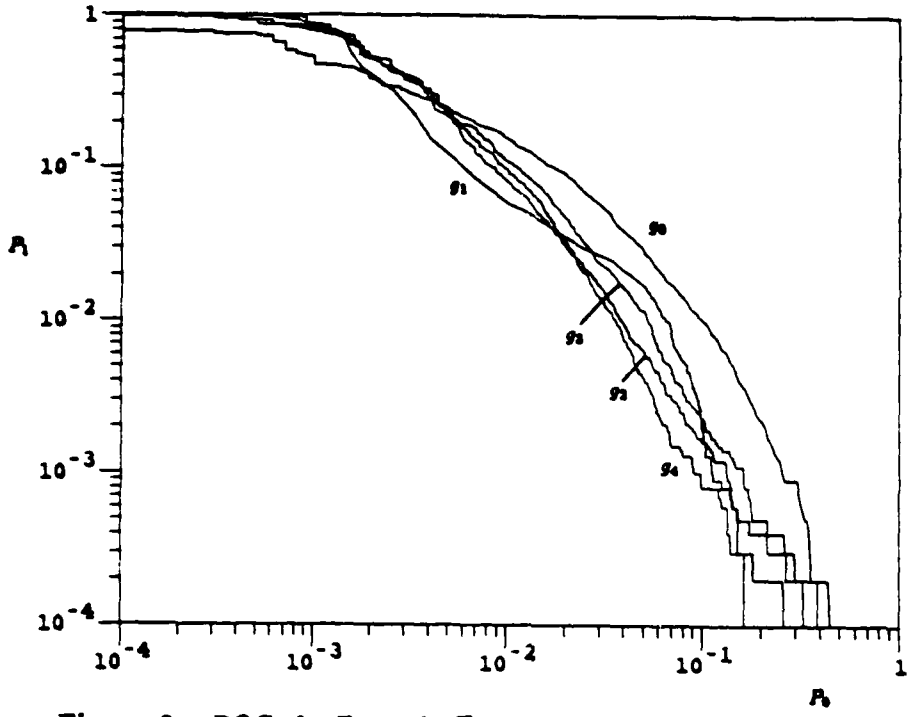

Figure 9. ROCs for Example E_3.

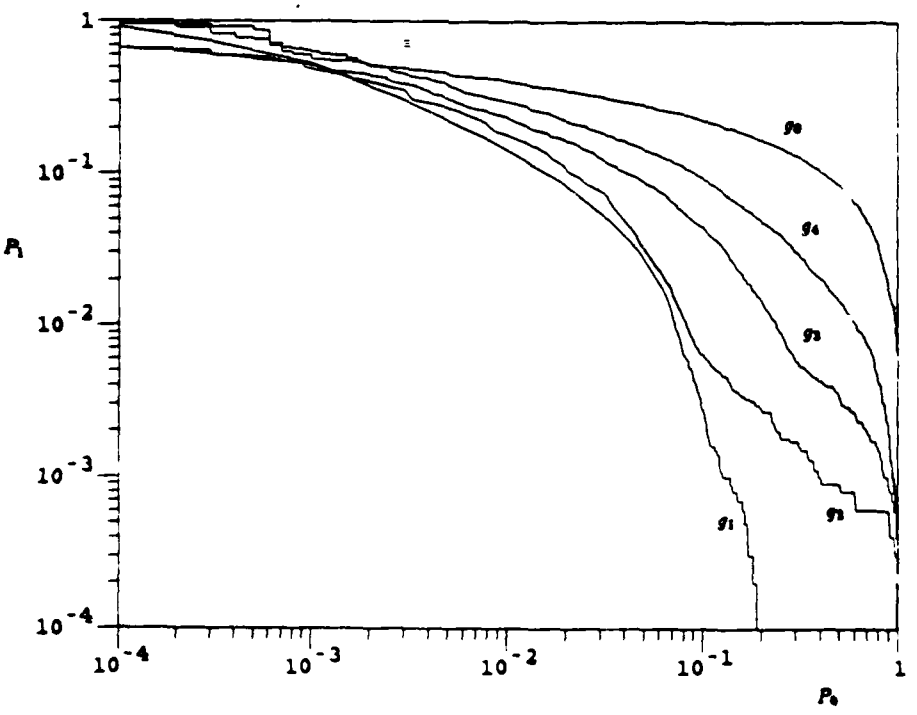

Figure 10. ROCs for Example E_4.

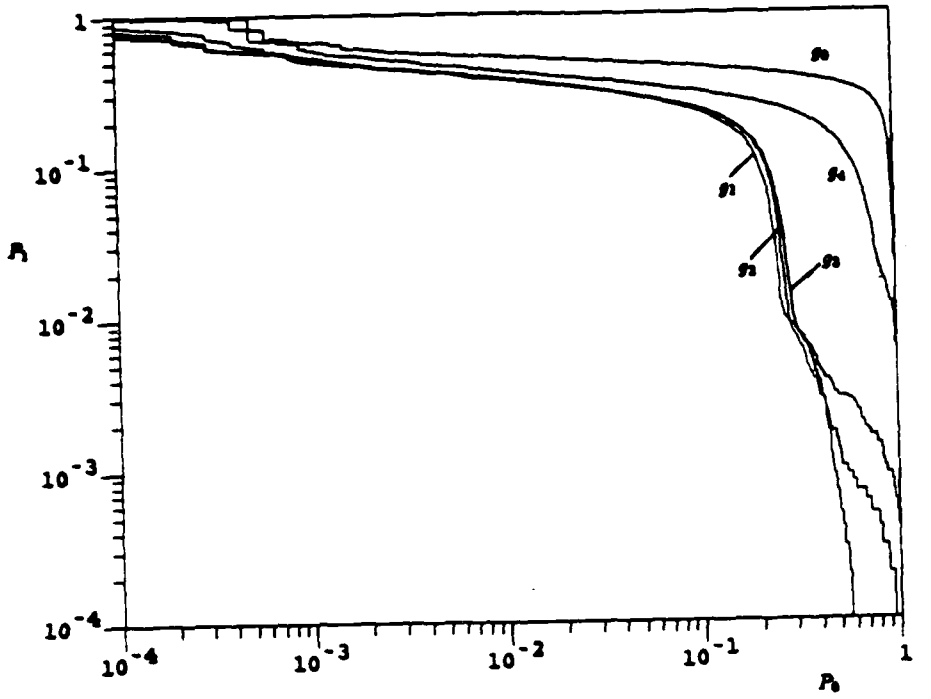

Figure 11. ROCs for Example E_5.

our discussion concerning the results from E_2 apply here, and we may interpret this as a case in which the performance from E_2 is degraded due to the increased dependency under H_0.

In Figure 11 we observe the results for E_5, where we hope to discern the change in the performance from E_2 by taking a smaller sample size. We find here that the performance of each nonlinearity has declined, as is to be expected. In the minimax region the ordering in essentially the same as that in E_2, although there is a small region where g_2 performs best. Clearly, though, the performance of g_1, g_2, and g_3 are practically the same.

5.4 Conclusion

In conclusion, we may wish to consider again the questions which were posed in the beginning of this chapter. First, regarding the validity of the performance measures, the simulation results and the values of the performance measures do not necessarily correlate well. In other words, knowing the values of a particular performance measure for two given nonlinearities, we may not be able to predict which nonlinearity will perform better in the simulations. Indeed, the nonlinearity g_4 does not maximize any of the performance measures S_i yet it has shown the best performance in the example E_1. This does not mean that the performance measures are not useful. On the contrary, they provide us with a method for calculating other nonlinearities which, as evidenced here, might prove to perform better than the iid nonlinearity. Nor does this mean that the theory is flawed, since the tests used here required finite sample sizes, an aspect which was neglected in the theory. One consequence of the finite sample size is that the distributions of the test statistics are not truly Gaussian. In fact, the distributions of the test statistics T_0 and T_1 are strongly skewed. Although this might be undesirable from the standpoint of the theory alone, this phenomenon is not necessarily undesirable in practice. Consider the test statistic T_1, for example. The variance of T_1 under the hypothesis H_0 is extremely large compared to the variance under H_1 and the compared to the difference between the means under the two hypotheses. However, because the distribution of T_1 is skewed to the left, most of the outliers under H_0 fall away from the threshold, and this results in a generally good performance.

Second, concerning the performance of memoryless discriminators for dependent processes, is has been demonstrated that in a situation of weak dependency, the iid

discriminator can be difficult to improve upon, while for a situation of strong dependency, the methods derived here show definite improvement over the iid discriminator. We might now conjecture as to how such an improvement might come about. If the dependency is strong under one of the hypotheses, say H_1, then the result of maximizing S_1 will likely lead to an improvement. This is because maximizing S_1 will result in a small value of σ_1^2, the effect of which is to minimize the much stronger dependency condition under H_1. This conjecture seems to be corroborated by the simulation results from E_2, E_4, and E_5 where the dependency under H_1 is stronger than that under H_0. If, however, there is strong dependency under both of the hypotheses, then perhaps the best approach would be to maximize S_2 or S_3, since in this way both of the dependency conditions are minimized. The relevant example here is E_4, where g_1 still performs best. However, as we noted above, the dependency is still somewhat stronger under H_1 than under H_0. The basic premise of the method described here is that when using memoryless discriminators for dependent processes, the best one can do to deal with the dependency conditions is to minimize their effects. The performance measures S_i provide framework for doing so.

It is satisfying to observe in the simulation results that g_2 and g_3 perform comparably, though g_2 generally performs slightly better. This is important because the calculation of g_3 is much easier than the calculation of g_2. Considering the good overall performance of the nonlinearity g_3, especially when there is strong dependency under both hypotheses, and the relatively simple calculation needed to determine it, this nonlinearity might be preferred in most situations.

The theory which has been presented here is entirely based on central limit theory and the assumption of large sample sizes. Indeed, all of the performance measures derived

here have been asymptotic performance measures. Admittedly, this approach might seem to be too narrow in its application to be useful in situations where a relatively moderate or small samle size is required. However, when faced with the problem of discriminating between two possible sources for an observed random process in which there is correlation in the observations, there is little else that one can do other than the LRT. Therefore, the work here is significant in that it does present an alternative to the LRT: the memoryless decision rule. In situations where the LRT cannot be implemented and there is decorrelation of the observations with time, one might be tempted to assume that the observations are iid, thereby leading to a memoryless discriminator. We have seen here several alternative memoryless discriminators, some of which might improve upon the iid discriminator. Furthermore, the simulation results have demonstrated that good results can be obtained for even moderate sample sizes. One possible area for future research might be to examine more thoroughly the performance for various sample sizes, and we might also note that these memoryless discriminators are ideally suited for sequential discrimination. We have also seen how some of the results can be made robust. Robust discrimination is important in many applications where circumstances might vary from test to test, such as a situation of radar discrimination of targets. Thus these robustness results are also significant, and the application of these results to radar problems might also be an area for future research.

APPENDIX A

The Rayleigh Distribution

Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be independent and identically distributed Gaussian random vectors such that $\mathrm{E}\, X_i = \mathrm{E}\, Y_i = 0$ and $\mathrm{Var}\, X_i = \mathrm{Var}\, Y_i = \theta_i$ for $i = 1, \ldots, n$. Suppose also that $\mathrm{E}\, X_i X_j / \sqrt{\theta_i \theta_j} = \mathrm{E}\, Y_i Y_j / \sqrt{\theta_i \theta_j} = \rho_{ij}$. Then the random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ defined by $Z_i = \sqrt{X_i^2 + Y_i^2}$ has a *Rayleigh* distribution. For $n > 2$, the n-dimensional density involves $n - 1$ iterated integrations and does not have a closed form expression. The bivariate density for (Z_i, Z_j) is

$$f_{Z_i Z_j}(u, v) = \frac{uv}{(1 - \rho_{ij}^2)\theta_i \theta_j} \exp\left\{ -\frac{1}{2(1 - \rho_{ij}^2)} \left[\frac{u^2}{\theta_i} + \frac{v^2}{\theta_j} \right] \right\} I_0 \left[\frac{\rho_{ij} uv}{(1 - \rho_{ij}^2)\sqrt{\theta_i \theta_j}} \right], \quad (u \geq 0, v \geq 0) \qquad (A1)$$

where I_0 is the modified Bessel function of the first kind of order 0. If the vectors $\mathbf{X}$ and $\mathbf{Y}$ are stationary, then $\mathbf{Z}$ is stationary and the density for (Z_1, Z_{j+1}) can be written

$$f_{Z_i Z_j}(u, v) = \frac{uv}{(1 - \rho_j^2)\theta^2} \exp\left[-\frac{u^2 + v^2}{2(1 - \rho_j^2)\theta} \right] I_0 \left[\frac{\rho_j uv}{(1 - \rho_j^2)\theta} \right], \quad (u \geq 0, v \geq 0) \qquad (A2)$$

where $\theta = \theta_1 = \theta_{j+1}$ and $\rho_j = \rho_{1,j+1}$. The marginal density takes the form

$$f(u) = \frac{u}{\theta} \exp\left(-\frac{u^2}{2\theta} \right), \qquad (u \geq 0). \qquad (A3)$$

The moments of the Rayleigh random variable Z are given by

$$\mathrm{E}\, Z^n = (2\theta)^{n/2} \Gamma\left(\frac{n}{2} + 1 \right) \qquad (n = 1, 2, \ldots) \qquad (A4)$$

where Γ denotes the gamma function.

APPENDIX B

The Lognormal Distribution

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a Gaussian random vector with $\mathrm{E}\, X_i = \lambda_1^{(i)}$ and $\mathrm{Var}\, X_i = \lambda_2^{(i)}$. Suppose also that $\mathrm{E}\, X_i X_j / \sqrt{\theta_i \theta_j} = \mathrm{E}\, Y_i Y_j / \sqrt{\theta_i \theta_j} = \rho_{ij}$. Then the random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ defined by $Z_i = \exp(X_i)$ has a *lognormal* distribution. If F_X and F_Z denote the n-dimensional distribution functions for the Gaussian random variables and lognormal random variables, respectively, then we have the relation

$$F_Z(z_1, \ldots, z_n) = F_X(\log z_1, \ldots, \log z_n). \tag{B1}$$

We may therefore obtain a relation for the densities by differentiating, and this yields

$$f_Z(z_1, \ldots, z_n) = \frac{1}{z_1 \cdots z_n} f_X(\log z_1, \ldots, \log z_n) \tag{B2}$$

The explicit expression for the bivariate density for (Z_i, Z_j) is

$$f_{Z_i Z_j}(u, v) = \left[2\pi uv \sqrt{\lambda_2^{(i)} \lambda_2^{(j)} (1 - \rho_{ij}^2)}\right]^{-1} \times$$

$$\exp\left\{ -\frac{1}{2(1-\rho_{ij}^2)} \left[\frac{(\log u - \lambda_1^{(i)})^2}{\lambda_2^{(i)}} - \frac{2\rho_{ij}(\log u - \lambda_1^{(i)})(\log v - \lambda_1^{(j)})}{\sqrt{\lambda_2^{(i)} \lambda_2^{(j)}}} + \frac{(\log v - \lambda_1^{(j)})^2}{\lambda_2^{(j)}} \right] \right\} \tag{B3}$$

If the random vector $\mathbf{X}$ is stationary, then so is the vector $\mathbf{Z}$, and the density for (Z_1, Z_{j+1}) can be written

$$f^j(u, v) = \left[2\pi uv \lambda_2 \sqrt{(1 - \rho_j^2)}\right]^{-1} \times$$

$$\exp\left\{ -\frac{(\log u - \lambda_1)^2 - 2\rho_j (\log u - \lambda_1)(\log v - \lambda_1) + (\log v - \lambda_1)^2}{2\lambda_2 (1 - \rho_j^2)} \right\} \tag{B4}$$

where $\lambda_1 = \lambda_1^{(1)} = \lambda_1^{(j+1)}$ and $\rho_j = \rho_{1,j+1}$. The marginal density has the form

$$f(u) = \frac{1}{u\sqrt{2\pi\lambda_2}} \exp\left\{-\frac{(\log u - \lambda_1)^2}{2\lambda_2}\right\}. \qquad (B5)$$

The moments of the lognormal random variable Z are given by

$$\mathrm{E}\, Z^n = \exp\left[n\lambda_1 + \frac{1}{2}n^2\lambda_2\right] \qquad (n = 1, 2, \ldots). \qquad (B6)$$

REFERENCES

[1] H. V. Poor and J. B. Thomas, "Memoryless Discrete-Time Detection of a Constant Signal in m-Dependent Noise," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 54–61, Jan. 1979.

[2] D. R. Halverson and G. L. Wise, "Discrete-Time Detection in ϕ-Mixing Noise," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 189–198, Mar. 1980.

[3] J. S. Sadowsky and J. A. Bucklew, "A Nonlocal Approach for Asymptotic Memoryless Detection Theory," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 115–120, Jan. 1986.

[4] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968, pp. 166–195.

[5] J. F. Barrett and D. G. Lampard, "An Expansion For Some Second-Order Probability Distributions and Its Application to Noise Problems," *IRE Trans. Inform. Theory*, vol. IT-1, pp. 10–15, March 1955.

[6] G. E. Noether, "On a Theorem of Pitman," *Ann. Math Stat.*, vol. 26, pp.64–68, 1955.

[7] P. J. Huber, "A Robust Version of the Probability Ratio Test," *Ann. Math. Statist.* vol. 36, pp. 1753–1758.

[8] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part 1*, J. Wiley: New York, 1968.

[9] J. S. Sadowsky, "A Maximum Variance Model for Robust Detection and Estimation with Dependent Data," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 220–226, March 1986.

[10] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag: New York, 1988.

[11] J. H. Miller and J. B. Thomas, "Detectors for Discrete-Time Signals in Non-Gaussian Noise," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 241–250, March 1972.

[12] G. V. Moustakides and J. B. Thomas, "Optimum Detection of a Weak Signal with a Minimal Knowledge of Dependency," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 97–102, Jan. 1986.

[13] P. J. Huber, "Robust Estimation of a Location Parameter," *Ann. Math. Statist.* vol. 35, pp. 73–101.

[14] P. J. Huber and V. Strassen, "Minimax Tests and the Neyman-Pearson Lemma for Capacities," *Ann. Stat.*, vol. 1, pp. 251–263, 1973.

[15] H. V. Poor, "On Robust Wiener Filtering," *IEEE Trans. Auto. Control* vol. AC-25, pp. 531–536, June 1980.

[16] H. V. Poor, "Robust Decision Design Using a Distance Criterion," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 575–587, September 1980.

[17] S. A. Kassam, "Robust Hypothesis Testing For Bounded Classes of Probability Densities," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 242–247, 1981.

[18] K. S. Vastola and H. V. Poor, "On the p-point Uncertainty Class," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 374–376, 1984.

CURRICULUM VITAE

Name: Douglas Wayne Sauder

Address: 11003 McKay Road
Fort Washington, MD 20744.

Degree and date to be conferred: Master of Science, 1988.

Major: Electrical Engineering.

Date of birth: August 16, 1961.

Place of birth: Pekin, Illinois.

Secondary education: Independent Baptist Academy
Clinton, Maryland.

Collegiate institutions attended	*Dates*	*Degree*	*Date of Degree*
Tennessee Temple University	9/79–1/83		
University of Tennessee at Chattanooga	1/83–5/85	B.A.	May 1985
University of Maryland, College Park	9/85–5/88	M.S.	May 1988

Professional publications:

D. Sauder and E. Geraniotis, "Optimal Memoryless Discrimination of Mixing Processes," Proceedings of the Conference on Information Systems and Sciences, 1988.

Professional positions held:

Graduate teaching assistant	Mathematics Department University of Maryland College Park, MD.	9/85–1/87
Summer intern	Analytical Services (ANSER) Suite 800 Jefferson Davis Hwy. Arlington, VA.	6/86–8/86
Graduate Research Fellow (Naval Research Laboratory)	Systems Research Center University of Maryland College Park, MD.	1/87–5/88
Electrical Engineer	Locus 2560 Huntington Avenue Alexandria, VA. 22303	6/88–