THESIS REPORT
Ph.D.
Institute for Systems Research

Non-Gaussian Methods for Signal Discrimination

by D.W. Sauder
Advisor: E.A. Geraniotis

The Institute for Systems Research is supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), Industry and the University

Ph.D. 93-2

Abstract

Title of Dissertation: Non-Gaussian Methods for Signal Discrimination

Douglas Wayne Sauder, Doctor of Philosophy, 1992

Dissertation directed by: Evaggelos A. Geraniotis
Professor
Department of Electrical Engineering

This dissertation addresses the problem of deciding which of two random signals is being received after observing a finite number of observations of that signal. The methods presented are varied, each addressing a different aspect of the problem.

First, we consider discriminators that involve a test statistic formed as a sum of one-step memory nonlinearities. The optimal one-step memory nonlinearity is obtained as the solution of an integral equation so as to maximize a generalized signal-to-noise ratio. The optimal one-step memory nonlinearity for weak-signal detection is also discussed. We obtain a partial "robustness" result, in which a guarateed level of performance is proved. Our simulation results show that optimal one-step memory discriminators yield a significant improvement over optimal memoryless discriminators.

Second, we consider likelihood ratio test (LRT) discriminators based on non-Gaussian multivariate probability density functions (pdfs). We propose that multivariate pdfs be constructed from the χ^2 distribution with a memoryless transformation to shape the univariate marginal pdf to any specified univariate pdf. With the χ^2 random variables constructed from underlying Gaussian autoregressive processes, we show that the multivariate

χ^2 pdfs may be numerically computed. As an alternative to multivariate pdfs constructed as transformations of the multivariate Gaussian pdf, we show that the transformed χ^2 multivariate pdfs can yield much better LRT discriminators in some cases.

Third, we consider LRT discriminators based on multivariate pdfs constructed from specified univariate pdfs via a hidden Markov model (HMM). These are not the usual nonstationary HMMs; the HMMs we consider are stationary with fixed univariate marginals. Estimation methods for identifying the parameters of the HMMs are discussed in detail. Parsimonious models with a small number of parameters are also discussed. Simulation results show that the models can be used to construct good LRT discriminators.

Finally, we consider the problem of designing a worst-case signal for the signal discrimination problem. The problem is formulated as a game that pits a signal designer against a discriminator designer. The signal designer tries to deceive the discriminator by replicating a signal of interest to the discriminator designer. Power constraints on the spurious signal signal prevent it from exactly replicating the statistics of the genuine signal. The problem is solved for a lower bound on time-averaged power of the spurious signal and a lower bound on expected power of the spurious signal.

NON-GAUSSIAN METHODS FOR SIGNAL DISCRIMINATION

Douglas W. Sauder

Copyright c 1993. Douglas W. Sauder. All Rights Reserved.

NON-GAUSSIAN METHODS FOR SIGNAL DISCRIMINATION

by

Douglas Wayne Sauder

Dissertation submitted to the Faculty of the Graduate School
of The University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1992

Advisory Committee:

Professor Evaggelos A. Geraniotis, Advisor
Professor Lee D. Davisson
Professor Benjamin Kedem
Assistant Professor Thomas Fuja
Assistant Professor Adrian Papamarcou

Contents

List of Figures iii

List of Tables v

1 Introduction 1

2 Optimal Discriminators Under a Generalized Signal-to-Noise Ratio 4
- 2.1 Introduction . . . 4
- 2.2 Problem Formulation . . . 6
- 2.3 Derivation of Optimal Nonlinearity for Signal Discrimination . . . 10
- 2.4 Derivation of the Optimal Nonlinearity for Weak-Signal Detection . . . 14
- 2.5 A Lower Bound for Worst-Case Performance . . . 15
- 2.6 Simulation Studies . . . 18
  - 2.6.1 Discrimination Between Rayleigh and Lognormal Correlated Time-Series . . . 18
  - 2.6.2 Detection of a Weak-Signal in Correlated Cauchy Noise . . . 19
- 2.7 Discussion . . . 24

3 Models for Positive-Valued Random Processes 26
- 3.1 Introduction . . . 26
- 3.2 Preliminary Theory . . . 29
  - 3.2.1 Transformation Models . . . 29
  - 3.2.2 The χ^2 Distribution . . . 30
  - 3.2.3 Autoregressive Models . . . 37
- 3.3 Computation of the Probability Density Functions . . . 39
  - 3.3.1 First-Order Markov PDF . . . 39
  - 3.3.2 Monte Carlo Integration . . . 41
  - 3.3.3 Gaussian Quadrature . . . 43
  - 3.3.4 Examples . . . 46
- 3.4 Parameter Estimation . . . 49
- 3.5 A Comparison of the PDFs . . . 55
- 3.6 Simulation Results . . . 59
- 3.7 Discussion . . . 67

4 Hidden Markov Models **68**
4.1 Introduction 68
4.2 Derivation of the Multivariate Distributions 69
4.3 Computation of the HMM/OS Multivariate Densities 75
4.4 Examples 80
4.5 Parameter Estimation 82
4.5.1 Least Squares Estimation 82
4.5.2 Minimum Discrimination Information Estimation 93
4.5.3 Maximum Likelihood Estimation 95
4.6 Simulation Studies 97
4.7 Discussion 101

5 Signal Discrimination Games **103**
5.1 Introduction 103
5.2 Problem Formulation 106
5.2.1 Players and Strategies 106
5.2.2 The Game Payoff 109
5.3 Discrimination Game with a Lower Bound on Time-Averaged Power 112
5.3.1 The Optimal Strategies 112
5.3.2 Case of I.I.D. Observations 113
5.3.3 Method for Generating the Spurious Signal 114
5.4 Discrimination Game with a Lower Bound on Expected Power 115
5.4.1 Sufficient and Necessary Conditions for the Maximin Distribution . 116
5.4.2 Interpretation of the Necessary Conditions 127
5.5 Simulation Studies 132
5.6 Discussion 135

6 Conclusion **137**

A Rayleigh and Lognormal Probability Density Functions **139**

B Cauchy Probability Density Functions **141**

Bibliography **143**

List of Figures

2.1 Graph over the region $[0,16] \times [0,16]$ of the optimal one-step memory nonlinearity for discrimination of Rayleigh vs. lognormal processes. 20
2.2 Plots of the error probabilities for the i.i.d. discriminator $T^{(0)}$, the optimal memoryless discriminator $T^{(1)}$, and the optimal one-step memory discriminator $T^{(2)}$ for a sample size of $n = 25$. 21
2.3 Plots of the error probabilities for the i.i.d. discriminator $T^{(0)}$, the optimal memoryless discriminator $T^{(1)}$, and the optimal one-step memory discriminator $T^{(2)}$ for a sample size of $n = 50$. 22
2.4 Graph over the region $[-233,233] \times [-233,233]$ of the optimal one-step memory nonlinearity for signal detection in Cauchy noise with $\tau = 1$. 23
2.5 Graph over the region $[-233,233] \times [-233,233]$ of the optimal one-step memory nonlinearity for signal detection in Cauchy noise with $\tau = 15$. 23

3.1 Contour plot of bivariate uniform pdf obtained by transformation of a $\chi^2(1)$ distribution. 56
3.2 Contour plot of bivariate uniform pdf obtained by transformation of a $\chi^2(2)$ distribution. 57
3.3 Contour plot of bivariate uniform pdf obtained by transformation of a $\chi^2(3)$ distribution. 57
3.4 Contour plot of bivariate uniform pdf obtained by transformation of a $\chi^2(4)$ distribution. 58
3.5 Contour plot of bivariate uniform pdf obtained by transformation of a Gaussian distribution. 58
3.6 Contour plot of bivariate exponential pdf obtained by transformation of a $\chi^2(1)$ distribution. 60
3.7 Contour plot of bivariate exponential pdf obtained by transformation of a $\chi^2(2)$ distribution. 60
3.8 Contour plot of bivariate exponential pdf obtained by transformation of a Gaussian distribution. 61
3.9 Plot of the error probabilities for lognormal vs. Rayleigh distributions, based on LRT for transformed Gaussian and transformed $\chi^2(2)$. Sample size is $n = 10$. 62
3.10 Plot of the error probabilities for LRT for exponential random variables with different AR order k. Sample size is $n = 20$. Number of nodes if $N = 20$. . . 63

3.11 Plot of the error probabilities for LRT for exponential random variables with different AR order k. Sample size is $n = 40$. Number of nodes if $N = 20$. . . 65
3.12 Plot of the error probabilities for LRT for exponential random variables with different AR order k. Sample size is $n = 40$. Number of nodes is $N = 10$. . 66

4.1 Discriminator error probabilities for the LRT discriminator with $N = 50$ samples per test. 101
4.2 Discriminator error probabilities for the LRT discriminator with $N = 100$ samples per test. 102

5.1 Plots of three different receiver operating characteristics (ROCs) 111
5.2 Example of a situation where the minimum Bayes risk occurs at a point on the ROC where $P_0 = 0$ 122
5.3 Plot of the probability density function f_0. 128
5.4 Plot of the general form of the maximin density f_1. 129
5.5 Plot of the maximin density f_1 when $u = u_{\min}$. 129
5.6 Plot of the maximin density f_1 when $u = u_{\max}$. 130
5.7 ROC for f_1 when $u = u_{\min}$ and $u = u_{\max}$. 131
5.8 Plot of the common value of the error probabilities vs. average power for i.i.d. signals. 133
5.9 Plot of the common value of the error probabilities vs. average power for correlated signals. 134
5.10 Plot of the common value of the error probabilities vs. average power for i.i.d. signals. 135

List of Tables

2.1 Efficacies of various detectors vs. correlation strength τ. 20
2.2 ARE of various detectors relative to the i.i.d. locally optimal detector vs. correlation strength τ. 21

3.1 Values of the computed pdf for $k = 1$ and $d = 2$. 47
3.2 Values of the computed pdf for $k = 1$ and $d = 3$. 47
3.3 Logarithm of the computed pdf for $k = 1$ and $d = 2$. 48
3.4 Logarithm of the computed pdf for $k = 1$ and $d = 3$. 48
3.5 Relative error of the log pdf for $k = 1$ and $d = 2$. 48
3.6 Relative error of the log pdf for $k = 1$ and $d = 3$. 48
3.7 Values of the computed pdf for $k = 2$ and $d = 2$. 49
3.8 Values of the computed pdf for $k = 2$ and $d = 3$. 49
3.9 Logarithm of the computed pdf for $k = 2$ and $d = 2$. 50
3.10 Logarithm of the computed pdf for $k = 2$ and $d = 3$. 50
3.11 Relative error of the log pdf for $k = 2$ and $d = 2$. 50
3.12 Relative error of the log pdf for $k = 2$ and $d = 3$. 50

4.1 Results of simulations for the esimation methods 99

Chapter 1

Introduction

A classic problem in radar signal processing for many years has been the problem of signal discrimination. Often referred to in the context of military radar systems as the problem of target identification, this problem involves observing a radar signal and deciding, on the basis of the observed signal, on the signal's source. In some cases, the source may be entirely noise—no object is present. In other cases, the source may be electromagnetic energy scattered by an object that intercepted the radar's transmitted signal in addition to noise. If indeed an object is present, then the problem is to identify the object from among several alternatives.

Although much is known about the scattering and propagation of radar signals, a deterministic approach to the signal discrimination problem seems at present to be a nearly impossible task. The interaction of the various elements that combine to form the radar signal are just too complex. Consequently, the standard approach is a probabilistic (or statistical) approach. Radar signals are modeled as stochastic processes, and the methods employed in the signal discrimination problem are statistical in nature.

This dissertation presents some new techniques for the discrimination of random signals. While the methods are broadly applicable to problems in diverse fields, the problem of target identification in radar signal processing has been the primary motivation and serves as a unifiying element. The basic problem that is common to all of the techniques presented here is to correctly identify a stationary random signal from observations of that signal. The different techniques we consider apply to different facets of the problem. The fact that diverse techniques may all be applied to the same basic problem attests the problem's complexity. We are far from obtaining the ultimate solution.

The signal discrimination problem is a mature one, and a lot of work has been done in this area. In recent years there have been some new trends in statistical techniques that have been applied to this problem. Notable among these are large deviations theory, higher-order statistics, and nonparametric methods such as neural networks and learning vector quantization (LVQ).

Large deviations theory pertains to the rate of convergence to zero of so-called "rare events" as the number of samples taken increases without bound [5]. In the context of signal discrimination, the rare events are incorrect decisions by the discriminator. Thus the large deviations methods are concerned with the rate at which the error probabilities converge to zero as the number of samples taken per test increases [52]. An implicit assumption required in the application of the large deviations results is that the sample size for discrimination is relatively large.

Higher-order cumulants have been hailed for their ability to defeat additive Gaussian noise with unknown autocovariance. The first- and second-order cumulants are the mean and autocovariance functions of a process. The higher-order cumulants, then, are third-order and higher. It is a well-known property that the third-order cumulants of symmetrically distributed random variables are zero, and that all of the higher-order cumulants of Gaussian random variables are zero. This fact accounts for the ability, at least in principle, for discrimination techniques based on higher-order cumulants to defeat additive Gaussian noise [15] [20]. In practice, higher-order sample cumulants can have relatively large variances, requiring, again, large sample sizes.

As neural networks have become quite popular in recent years, their proponents have not left the signal discrimination problem unscathed. Neural networks have been successfully used for discrimination in a variety of fields, including speech recognition, radar and sonar signal classification [17], and other areas. These successes stand, despite the fact that the popular back-propagation algorithm for training the networks has been shown to be inadequate in some cases [3]. Another nonparametric approach to the discrimination problem lies in learning vector quantization (LVQ) [31]. LVQ is not unrelated to neural networks; LVQ algorithms are often implemented with neural network architectures.

For the most part, the research conducted for this thesis has steered clear of the trends just mentioned. There are primarily two reasons for this. The first reason is that the focus has been on the problem of discrimination of radar signals when it is known that an object has been detected. It seems, however, that the vast majority of the detection

literature focuses on the problem of detecting the presence of a signal in additive noise. Thus the basic problem is not to determine which from among several objects has been detected, but rather to determine whether an object has been detected or whether the signal is pure noise. This is the so-called weak-signal detection problem. The trends mentioned above seem to be more suitable for this weak-signal problem. The second reason is that an attempt was made to be practical, and asymptotic techniques that require a large sample size have mostly been avoided. The topic considered in Chapter 5 on signal detection games, in particular, has been motivated by a practical problem.

The results that are presented are diverse, yet they address different facets of the basic signal detection problem. In Chapter 2, simplicity of the discriminator implementation is a major concern, as we consider discriminators that involve a one-step memory nonlinearity. The nonlinearity is used to efficiently compute a test statistic that is used in a threshold test. In Chapters 3 and 4, the sample size is the major issue. Here we look at discriminators that are specifically designed for use with small or medium sample sizes. The idea is to stay away from asymptotic techniques, such as large deviations or the central limit theorem. The results depend upon the successful fitting of a model to data, in order that a likelihood ratio test may be performed. This approach is commonly used in maximum likelihood pattern classification. In Chapter 3, we specifically consider models that are appropriate for positive-valued random processes. These results are especially useful for modeling the envelope of a complex signal, as is common in radar signal processing. The hidden Markov models (HMMs) in Chapter 4 represent a new twist to an old idea. We consider HMMs in which the univariate marginal distributions are explicitly specified. In a sense, these models can be considered a blend of non-parametric and parameteric models: the univariate distributions are parametric, while the dependency structure that is seen in the multivariate distributions has potentially a large number of parameters, as is characteristic of nonparametric models. Chapter 5 stands somewhat apart from the other chapters, but nevertheless addresses an important practical aspect of the discrimination problem, namely, the designing of a decoy signal. This topic is related to the area of robustness, but differs in that the emphasis is on the performance from the point of view of the spurious signal designer, rather than the point of view of the discriminator. These results are important to the discriminator designer, however, as they enable him to understand how a spurious signal would be designed.

Chapter 2

Optimal Discriminators Under a Generalized Signal-to-Noise Ratio

2.1 Introduction

It is not often that one can obtain a closed form expression for the error probabilities associated with a particular discriminator, the main reason being that the distribution of the test statistic is too intractable to work with. Consequently, researchers have devised alternative performance measures. A very well-known such performance measure is the signal-to-noise ratio (SNR), which is used to measure the performance of a detector, or *filter*, for detecting the presence of a signal in additive noise. The SNR measures the ratio of a signal's strength (amplitude) to its statistical variability (variance). For problems that cannot be appropriately modeled as a signal in additive noise—problems in target identification, for example—the SNR is not appropriate. However, the idea of a performance measure comprising a measure of signal separation divided by a measure of statistical variability is still appealing. The performance measure that we utilize in this section is a performance measure of this type, and, consequently, we refer to it as a generalized SNR.

Although an SNR-type performance measure may be used with any test statistic that has computable first and second moments, one test statistic in particular, the test statistic based on a memoryless nonlinearity, has been especially popular for such analysis. This test statistic, which has been the focus of much research, is computed as the sum of observations taken at the output of a memoryless nonlinearity [see (2.3)]. The important

feature that makes it useful is the fact that the central limit theorem may be applied to obtain its asymptotic distribution. The central limit theorem is especially important in justifying the use of the efficacy performance measure.

Early papers in this area include Noether's paper [44], which introduced the notion of asymptotic relative efficiency (ARE) and the efficacy functional, a result now commonly referred to as the Pitman-Noether theorem [46]. For i.i.d processes, the nonlinearity that maximizes the efficacy performance measure is known to be the locally optimal nonlinearity [7]. Subsequent papers continued on the theme of the efficacy performance measure and the locally optimal nonlinearity. Miller and Thomas [40] considered nonconstant signals in non-Gaussian white noise. Poor and Thomas [47] continued the effort by accommodating m-dependent noise, and Halverson and Wise [19] further extended the results to the case of ϕ-mixing noise. While these results all involved the efficacy performance measure, a measure of local optimality, Sadowsky and Bucklew [52] started a new direction by deriving a nonlocal performance measure. Their work was continued by Sauder and Geraniotis [53], who compared several related nonlocal performance measures and addressed the issue of robustness.

The above references are related in that they all consider discriminators/detectors based on a memoryless nonlinearity. Although Halverson and Wise [19] mentioned the extension of their results to nonlinearities with memory, they did not go into detail. The work presented in this chapter begins at this point—an extensive body of literature on memoryless nonlinearities—but goes further in discriminator complexity by considering discriminators based on one-step memory nonlinearities.

We show in this chapter that one-step memory nonlinearities maximizing the generalized SNR (Section 2.3) and efficacy (Section 2.4) performance measures may be obtained through the solution of certain linear integral equations. For m-dependent observations, the integral equations yield precisely the optimal nonlinearities, while for ρ-mixing observations the optimal nonlinearities may be approximated as closely as desired by the integral equation solutions. Because the integral equations generally lack closed form solutions, numerical approximations are required, and we show how the approximations may be obtained by solving a large but manageable system of linear equations. We have also obtained discriminators that have a guaranteed level of performance over uncertainty classes for the distributions (Section 2.5). This result is related to minimax robustness, the difference being that we have not been able to prove that there actually exists a pair of distributions in the uncertainty

classes for which the lower bound on the performance is actually obtained. In our numerical results in Section 2.6, we compare the performance of several discriminators (i.i.d. LRT, optimal memoryless, and optimal one-step memory) for the problem of discriminating between Rayleigh and lognormal processes. We also show the ARE of several detectors relative to the i.i.d. (locally optimal) detector for the problem of signal detection in Cauchy noise. The results indicate the significant advantage in the discriminator or weak-signal detector performance that becomes possible through the use of one-step memory nonlinearities.

2.2 Problem Formulation

There are two related hypothesis testing problems that we consider here. The first of these is the signal discrimination problem, which involves deciding between the two hypotheses

$$\begin{aligned} H_0 &: \ \{X_k\} \text{ has the distribution } F_0 \\ H_1 &: \ \{X_k\} \text{ has the distribution } F_1 \end{aligned} \tag{2.1}$$

where we use F_i to denote symbolically all the known or assumed information about the distribution of process $\{X_k\}$ under H_i. We assume the process $\{X_k\}$ to be stationary under both hypotheses. Problems of this type arise in areas such as radar target discrimination, where one must decide which of two possible targets is present on the basis of returned radar pulses. In some instances, the distribution of a process is known in principle, as the underlying mechanism which generates the process is known, but a closed form expression for the complete multivariate density is not available (e.g. the Rice process which is obtained as the envelope of a complex Gaussian process with nonzero mean). For the derivation of the optimal one-step memory nonlinearities in this paper, all of the marginal probability densities of dimensions one, two, three, and four are required. Therefore, for this paper F_i must denote at least the complete set of marginal densities up to four dimensions. By contrast, only univariate and bivariate densities are required for the derivation of optimal memoryless nonlinearities [53].

A problem that is related to the discrimination problem is the weak-signal detection problem,

$$\begin{aligned} H_0 &: \ X_k = Y_k \\ H_1 &: \ X_k = Y_k + \theta, \end{aligned} \tag{2.2}$$

where $\theta > 0$ represents the constant signal and $\{Y_k\}$ the additive noise. The statistical

distribution F_0 of the noise is assumed to be known, in the sense described in the preceding paragraph. To obtain the optimal one-step memory nonlinearity derived here, the marginal densities corresponding to F_0 up to four dimensions must be available. So far, research in the weak-signal detection area has yielded memoryless nonlinearities that are optimal under the asymptotic relative efficiency (ARE) performance measure [47], [19] and robust [51]. These results require knowledge of the univariate and bivariate densities only.

The mechanism we employ to decide between H_0 and H_1 on the basis of a collection $\mathbf{x} = (x_1, \ldots, x_n)$ of observations involves comparing a test statistic T_n to a threshold γ. We decide H_1 if $T_n > \gamma$, decide H_0 if $T_n < \gamma$, and decide on some other basis if $T_n = \gamma$. Over the last decade, a considerable amount of attention has been devoted to the study of detectors and discriminators based on memoryless nonlinearities. These tests use a test statistic of the form

$$T_n(\mathbf{x}) = \sum_{i=1}^{n} g(x_i). \tag{2.3}$$

where g is a *memoryless nonlinearity*. This test statistic is appealing, when its performance is acceptable, because its simple form makes it easy to implement. If the observations are i.i.d. under both H_0 and H_1, then the likelihood ratio test has the structure required by (2.3); we take $g(x) = \log\left[f_1(x)/f_0(x)\right]$. The memoryless test statistic lends itself well to asymptotic analysis, too, since a central limit theorem may often be applied to yield its asymptotic distribution.

The use of the memoryless test statistic for the signal discrimination problem was explored fairly completely in [53], including an interesting presentation of simulation results that compared memoryless discriminators obtained by maximizing different generalized SNRs with the i.i.d. discriminator. To gain insight into the nature of the present problem, it is worth pointing out two situations in which the i.i.d. discriminator yielded better performance than the other "optimal" memoryless discriminators. The two conditions are these: (1) when the correlation is very weak and (2) when the univariate marginal densities f_1 and f_0 are significantly different. A plausible explanation is obvious. In the first case, the processes are so close to being i.i.d. that the i.i.d. discriminator, which is by definition the optimal discriminator in terms of the error probabilities when the observations are truly i.i.d., performs approximately optimally (in terms of the error probabilities). The other discriminators, which are optimal under a generalized SNR, cannot then compete. In the second case, the signals under the two hypotheses are so different that any SNR-type

performance measure is not likely to be useful in predicting of how discriminators perform under a criterion involving the error probabilities (e.g. Bayes risk). This conclusion was also reached by Sadowsky and Bucklew in [52]. Thus the optimal memoryless discriminators derived in [53] seem to offer the greatest improvement over the i.i.d. discriminator when the univariate densities are not too dissimilar and the correlation is moderate to strong. Intuition suggests, however, that in this case most of the discrimination potential lies in the correlation, especially if the strength of the correlation is significantly different under the two hypotheses. It is primarily this line of thinking which motivates us to investigate one-step memory nonlinearities for discrimination and detection. Because the memoryless test statistic (2.3) is our starting point, we shall refer to it often as a standard for comparison and discussion concerning the one-step memory test statistic.

The form of the test statistic we are most interested in here is

$$T_n(\mathbf{x}) = \sum_{j=1}^{n-1} g(x_j, x_{j+1}) \tag{2.4}$$

where $g(\cdot,\cdot)$ is a *one-step memory nonlinearity* that characterizes the test. Since we are concerned primarily about asymptotic performance, we ignore the possibility of an initial term $g_0(x_1)$. A decision is made by comparing T_n to a threshold, as in the case of the memoryless discriminator. The test statistic (2.4) maintains some of the advantages of the memoryless test statistic (2.3), while introducing a means to discriminate on the basis of correlation. In particular, with the proper condition of asymptotic independence of the observations, T_n obeys a central limit theorem. One must use caution, though, not to infer too much from the asymptotic normality of the test statistic. For example, if the error probabilities vanish asymptotically, as they do for a consistent test, one cannot use the central limit theorem to accurately approximate them. See [52] for a more complete discussion of issues related to nonlocal detection and an approach based on large deviations. The asymptotic normality of the test statistic is important, however, in the context of weak-signal detection when the efficacy performance measure (defined below) is used.

Dependency conditions that are sufficient for a central limit theorem (CLT) to hold are well known. For a good survey paper on CLTs for dependent data, see [45]. The two conditions we consider most useful are m-dependence and ρ-mixing. To define these conditions, let $\mathcal{F}_a^b$ denote the σ-field generated by the random variables $\{X_a, \ldots, X_b\}$. The process $\{X_k\}$ is said to be m-dependent under a distribution F if $\mathcal{F}_{-\infty}^k$ and $\mathcal{F}_{k+m}^{\infty}$ are independent for any k. If U and V are random variables which are measurable with respect

to $\mathcal{F}_{-\infty}^{k}$ and $\mathcal{F}_{k+n}^{\infty}$, respectively, and square integrable under the distribution F, then ρ_n is defined as $\sup_{U,V} \mathrm{Cov}(U,V)/\sqrt{\mathrm{Var}U\,\mathrm{Var}V}$ where the supremum is taken over all such U and V. The process $\{X_k\}$ is then defined to be ρ-mixing under the distribution F if the sequence $\{\rho_n\}$ converges to zero as $n \to \infty$. If the observation sequence $\{X_k\}$ is m-dependent, then the sequence $\{g(X_k, X_{k+1})\}$ is $(m+1)$-dependent for any square-integrable nonlinearity g. Also, if $\{X_k\}$ is ρ-mixing, so is $\{g(X_k, X_{k+1})\}$. Thus a central limit theorem applies as in the memoryless case. Here we assume that all processes are stationary and either m-dependent or ρ-mixing.

The asymptotic performance of the test statistic T_n may be measured by its asymptotic mean and variance, which have the expressions

$$\mu_i(g) = \lim_{n\to\infty} \frac{1}{n} \mathrm{E}_i\, T_n = \mathrm{E}_i\, g(X_1, X_2) \tag{2.5}$$

and

$$\begin{aligned} \sigma_i^2(g) &= \lim_{n\to\infty} \frac{1}{n} \mathrm{Var}_i\, T_n \\ &= \mathrm{Var}_i\, g(X_1, X_2) + 2\sum_{j=1}^{\infty} \mathrm{Cov}_i\,[g(X_1, X_2), g(X_{j+1}, X_{j+2})] \end{aligned} \tag{2.6}$$

where E_i, Var_i, and Cov_i respectively denote the expectation, variance, and covariance operators under H_i. When the process is m-dependent, then the upper limit on the sum in (2.6) is m.

In [53], the authors used for a performance measure a generalized signal-to-noise ratio

$$S_\zeta(g) = \frac{|\mu_1(g) - \mu_0(g)|^2}{(1-\zeta)\sigma_0^2(g) + \zeta\sigma_1^2(g)} \tag{2.7}$$

where $0 \le \zeta \le 1$. In this chapter, we take the optimal one-step memory nonlinearity to be the one which maximizes the performance measure S_ζ. The parameter ζ determines the weighting of the asymptotic variances. Although a value of ζ near $\frac{1}{2}$ yields a balanced mixture, we also obtained good simulation results in [53] for $\zeta = 1$, especially in the case of strong correlation under H_1 and weak correlation under H_0. It is also possible to take ζ to be a function of g. In particular, if ζ is taken to be

$$\zeta(g) = (2\sigma_0\sigma_1 + \sigma_1^2)/(\sigma_1^2 - \sigma_0^2), \tag{2.8}$$

then the performance measure becomes

$$S(g) = \frac{|\mu_1 - \mu_0|^2}{|\sigma_0 + \sigma_1|^2} \tag{2.9}$$

which is the performance measure obtained in [52]. In [53], it was explained how the optimal nonlinearity for (2.9) may be obtained by solving iteratively a series of linear integral equations. The idea is to choose an initial value for ζ, optimize the performance measure S_ζ, use the nonlinearity thus obtained to compute σ_0 and σ_1, and then compute a new value for ζ using (2.8). The procedure is then repeated until the value of ζ does not change significantly.

When the problem under consideration is weak-signal detection (2.2), the appropriate performance measure is the efficacy functional [46]

$$\eta(g) = \frac{|\mu_0'(g)|^2}{\sigma_0^2(g)} \tag{2.10}$$

where $\mu_0'(g) = (d/d\theta)\mathrm{E}_\theta\, g(X_1, X_2)\Big|_{\theta=0}$. We denote by E_θ the expectation taken under H_1 when the signal strength is θ. The use of the efficacy functional requires that

$$\lim_{\theta \to 0} \frac{\sigma_\theta^2(g)}{\sigma_0^2(g)} = 1 \tag{2.11}$$

where $\sigma_\theta^2(g)$ is the asymptotic variance of T_n when the signal strength is θ. A few other mild regularity conditions on the distribution are also required [46], including the uniform weak convergence of the distribution to normality for θ in an interval around zero. The relation of (2.10) to (2.7) is evident.

2.3 Derivation of Optimal Nonlinearity for Signal Discrimination

The one-step memory discriminator is characterized by the one-step memory nonlinearity $g(\cdot,\cdot)$ used in the test statistic T_n and the threshold γ. Once a nonlinearity g is given, the distribution of T_n can be estimated through simulation and the threshold determined on the basis of the estimated distribution and the desired error probabilities. Thus the important design step is the selection of the nonlinearity. Our task is to find the nonlinearity g which maximizes $S_\zeta(g)$.

To avoid problems with the interchanging of limits, we must begin with the optimization of $S_\zeta(g)$ for the m-dependent model. Results for the ρ-mixing model are obtained via approximation by m-dependent models with sufficiently large m. We first note that $S_\zeta(g)$ remains invariant under scaling of g; therefore we may maximize the numerator while

constraining the denominator to be constant. Introducing a Lagrange multiplier λ, our maximization problem becomes equivalent to maximizing the functional

$$J(g) = \mu_1(g) - \mu_0(g) - \lambda\Big[(1-\zeta)\sigma_0^2(g) + \zeta\sigma_1^2(g)\Big]. \tag{2.12}$$

This is the approach used in [47], [19], [52], [53]. A necessary condition for g to maximize J is that $\frac{d}{d\epsilon}J(g+\epsilon h)\Big|_{\epsilon=0} = 0$, for an arbitrary nonlinearity h satisfying $\mathrm{E}_i\left|h(X_1,X_2)\right|^2 < \infty$ for $i = 0, 1$. First we compute

$$\frac{d}{d\epsilon}\mu_i(g+\epsilon h)\Big|_{\epsilon=0} = \mathrm{E}_i\, h(X_1,X_2) \tag{2.13}$$

and

$$\begin{aligned}\frac{d}{d\epsilon}\sigma_i^2(g+\epsilon h)\Big|_{\epsilon=0} &= 2\mathrm{E}_i\, g(X_1,X_2)h(X_1,X_2) \\ &\quad + 2\sum_{j=1}^{m}\Big[\mathrm{E}_i\, g(X_1,X_2)h(X_{j+1},X_{j+2}) + \mathrm{E}_i\, g(X_{j+1},X_{j+2})h(X_1,X_2)\Big] \\ &\quad - 2(2m+1)\mathrm{E}_i\, g(X_1,X_2)\mathrm{E}_i\, h(X_1,X_2).\end{aligned} \tag{2.14}$$

We now introduce the notation $f_i^{(j,k)}$, $f_i^{(j,k,l)}$, and $f_i^{(j,k,l,p)}$ denoting the densities of (X_j,X_k), (X_j,X_k,X_l), and(X_j,X_k,X_l,X_p), respectively, under H_i. Using (2.13) and (2.14), we obtain

$$\begin{aligned}\frac{d}{d\epsilon}J(g+\epsilon h)\Big|_{\epsilon=0} &= \iint h(x,y)\Big[f_1^{(1,2)}(x,y) - f_0^{(1,2)}(x,y)\Big]\,dx\,dy \\ &\quad - 2\lambda\Big[(1-\zeta)W_0(g,h) + \zeta W_1(g,h)\Big]\end{aligned} \tag{2.15}$$

where the functional W_i is defined by

$$\begin{aligned}W_i(g,h) &= \iint g(x,y)h(x,y)f_i^{(1,2)}(x,y)\,dx\,dy \\ &\quad + \iiint\Big[g(x,y)h(y,u) + g(y,u)h(x,y)\Big]f_i^{(1,2,3)}(x,y,u)\,dx\,dy\,du \\ &\quad + \sum_{j=2}^{m}\iiint\Big[g(x,y)h(u,v) + g(u,v)h(x,y)\Big]f_i^{(1,2,j+1,j+2)}(x,y,u,v)\,dx\,dy\,du\,dv \\ &\quad - (2m+1)\iint g(x,y)f_i^{(1,2)}(x,y)\,dx\,dy\iint h(x,y)f_i^{(1,2)}(x,y)\,dx\,dy.\end{aligned} \tag{2.16}$$

The right-hand side of (2.15) becomes zero for an arbitrary h if and only if the part of the

integrand which multiplies h is identically 0. Thus we obtain the integral equation

$$\begin{aligned} f_1^{(1,2)}(x,y) - f_0^{(1,2)}(x,y) = 2\lambda g(x,y)\Big[(1-\zeta)f_0^{(1,2)}(x,y) + \zeta f_1^{(1,2)}(x,y)\Big] \\ + 2\lambda \int g(u,x)K_1(u,x,y)\,du + 2\lambda \int g(y,u)K_1(x,y,u)\,du \\ + 2\lambda \iint g(u,v)K_2(x,y,u,v)\,du\,dv \end{aligned} \tag{2.17}$$

where

$$K_1(u,x,y) = (1-\zeta)f_0^{(1,2,3)}(u,x,y) + \zeta f_1^{(1,2,3)}(u,x,y) \tag{2.18}$$

$$\begin{aligned} K_2(x,y,u,v) = \sum_{j=2}^{m}\Big[(1-\zeta)f_0^{(1,2,j+1,j+2)}(u,v,x,y) + \zeta f_1^{(1,2,j+1,j+2)}(u,v,x,y) \\ + (1-\zeta)f_0^{(1,2,j+1,j+2)}(x,y,u,v) + \zeta f_1^{(1,2,j+1,j+2)}(x,y,u,v)\Big] \\ - (2m+1)\Big[(1-\zeta)f_0^{(1,2)}(u,v)f_0^{(1,2)}(x,y) + \zeta f_1^{(1,2)}(u,v)f_1^{(1,2)}(x,y)\Big]. \end{aligned} \tag{2.19}$$

It is evident that λ determines the scaling of g, which is irrelevant to the performance, provided that the threshold is properly scaled. It seems most convenient to set $\lambda = \frac{1}{2}$, which yields the integral equation

$$\begin{aligned} f_1^{(1,2)}(x,y) - f_0^{(1,2)}(x,y) = g(x,y)\Big[(1-\zeta)f_0^{(1,2)}(x,y) + \zeta f_1^{(1,2)}(x,y)\Big] \\ + \int g(u,x)K_1(u,x,y)\,du + \int g(y,u)K_1(x,y,u)\,du \\ + \iint g(u,v)K_2(x,y,u,v)\,du\,dv. \end{aligned} \tag{2.20}$$

It is interesting to note that the generalized SNR (2.7) has a property similar to that of the SNR and the matched filter. If we multiply both sides of (2.20) by $g(x,y)$ and integrate, we find that, if g^* solves the integral equation, then (2.20) becomes

$$\mu_1(g^*) - \mu_0(g^*) = (1-\zeta)\sigma_0^2(g^*) + \zeta\sigma_1^2(g^*). \tag{2.21}$$

This gives us the value of the performance measure

$$S_\zeta(g^*) = \mu_1(g^*) - \mu_0(g^*) \tag{2.22}$$

evaluated at the optimal nonlinearity.

Under mismatch conditions, which is the case when the nonlinearity g is not the optimal nonlinearity for the densities $f_i^{(1,2)}$, $f_i^{(1,2,3)}$, and $f_i^{(1,2,j+1,j+2)}$ involved in (2.18)-(2.20), the performance measure $S_\zeta(g)$ can be computed from (2.7) where the numerator is

given by

$$\mu_1(g) - \mu_0(g) = \iint g(x,y)\left[f_1^{(1,2)}(x,y) - f_0^{(1,2)}(x,y)\right] dx\, dy \tag{2.23}$$

and the denominator by

$$\begin{aligned}(1-\zeta)\sigma_0^2(g) + \zeta\sigma_1^2(g) = &\iint g^2(x,y)\left[(1-\zeta)f_0^{(1,2)}(x,y) + \zeta f_1^{(1,2)}(x,y)\right] dx\, dy \\ &+ \iiint g(x,y)\left[g(u,x)K_1(u,x,y) + g(y,u)K_1(x,y,u)\right] du\, dx\, dy \\ &+ \iiiint g(x,y)g(u,v)K_2(x,y,u,v)\, dx\, dy\, du\, dv.\end{aligned} \tag{2.24}$$

The integral equation (2.20) generally cannot be solved for a closed form solution; numerical methods must be used. We recommend a normalization of (2.20), in order to facilitate a more accurate numerical solution, by introducing the substitution $h(x,y) = g(x,y)\sqrt{D(x,y)}$, with $D(x,y) = (1-\zeta)f_0^{(1,2)}(x,y) + \zeta f_1^{(1,2)}(x,y)$. The new integral equation

$$\begin{aligned}\frac{f_1^{(1,2)}(x,y) - f_0^{(1,2)}(x,y)}{\sqrt{D(x,y)}} = h(x,y) &+ \int \frac{h(u,x)K_1(u,x,y)}{\sqrt{D(u,x)D(x,y)}} du + \int \frac{h(y,u)K_1(x,y,u)}{\sqrt{D(y,u)D(x,y)}} du \\ &+ \iint \frac{h(u,v)K_2(x,y,u,v)}{\sqrt{D(u,v)D(x,y)}}\, du\, dv\end{aligned} \tag{2.25}$$

is more "balanced" and results in a more stable solution. In particular, the kernel $K_2(x,y,u,v)/\sqrt{D(u,v)D(x,y)}$ of the double integral is symmetric in the pairs (x,y) and (u,v).

To solve the integral equation numerically we introduce the nodes $\{x_0,\ldots,x_{N-1}\}$ and weights $\{w_0,\ldots,w_{N-1}\}$ from some numerical integration method (such as Simpson's rule). The numerical form of the integral equation (2.25) is then

$$\begin{aligned}\frac{f_1^{(1,2)}(x_i,x_j) - f_0^{(1,2)}(x_i,x_j)}{\sqrt{D(x_i,x_j)}} = u_{ij} &+ \sum_{k=0}^{N-1} \frac{u_{ki}K_1(x_k,x_i,x_j)w_k}{\sqrt{D(x_k,x_i)D(x_i,x_j)}} + \sum_{k=0}^{N-1} \frac{u_{jk}K_1(x_i,x_j,x_k)w_k}{\sqrt{D(x_i,x_j)D(x_j,x_k)}} \\ &+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} \frac{u_{kl}K_2(x_k,x_l,x_i,x_j)\, w_k\, w_l}{\sqrt{D(x_k,x_l)D(x_i,x_j)}}\end{aligned} \tag{2.26}$$

with $u_{i,j} \approx h(x_i,x_j) = g(x_i,x_j)\sqrt{D(x_i,x_j)}$. Eq. (2.26) can be solved as a system of N^2 linear equations in the N^2 variables u_{ij}, $i = 0,\ldots,N-1$, $j = 0,\ldots,N-1$. Note that, without the presence of the terms involving K_1, the matrix involved is symmetric.

The results of this section may be applied to ρ-mixing processes. The method involves approximating the ρ-mixing process by an m-dependent process; that is, one assumes that observations are independent if they are farther than m steps apart in time. The approximation is sufficient if m is sufficiently large. More precisely, if $g^{(m)}$ is the the nonlinearity obtained by solving the integral equation (2.20), then $S_\zeta(g^{(m)}) \to \sup_g S_\zeta(g)$, as $m \to \infty$. A proof of this result was given in [53] for memoryless nonlinearities; the extension to one-step memory nonlinearities is straightforward.

2.4 Derivation of the Optimal Nonlinearity for Weak-Signal Detection

For weak-signal detection, the hypotheses are given by (2.2), where $\{Y_i\}$ an m-dependent noise process and $\theta > 0$ is a constant signal. Thus, the n-variate densities under H_0 and H_1 are, respectively, $f^{(n)}(\mathbf{x})$ and $f^{(n)}(\mathbf{x} - \underline{\theta})$, where $\underline{\theta} = (\theta, \theta, \cdots, \theta)$. The performance measure here is the efficacy functional (2.10). The method for maximizing $\eta(g)$ is similar to that for maximizing $S_\zeta(g)$. Assuming that

$$\mu_0'(g) = \iint g(x,y) \frac{d}{d\theta} f^{(1,2)}(x-\theta, y-\theta)\Big|_{\theta=0} dx\, dy \tag{2.27}$$

as well as some other regularity conditions [47], [19], we can obtain the integral equation

$$\begin{aligned} -\left[\frac{\partial}{\partial x} f^{(1,2)}(x,y) + \frac{\partial}{\partial y} f^{(1,2)}(x,y)\right] &= g(x,y) f^{(1,2)}(x,y) \\ &\quad + \int \left[g(u,x) f^{(1,2,3)}(u,x,y) + g(y,u) f^{(1,2,3)}(x,y,u)\right] du \\ &\quad + \iint g(u,v) K(x,y,u,v)\, du\, dv \end{aligned} \tag{2.28}$$

where

$$\begin{aligned} K(x,y,u,v) = \sum_{j=2}^{m} &\left[f^{(1,2,j+1,j+2)}(u,v,x,y) + f^{(1,2,j+1,j+2)}(x,y,u,v)\right] \\ &- (2m+1) f^{(1,2)}(u,v) f^{(1,2)}(x,y) \end{aligned} \tag{2.29}$$

and $f^{(j,k)}$, $f^{(j,k,l)}$, and $f^{(j,k,l,p)}$ denote the densities of (Y_j, Y_k), (Y_j, Y_k, Y_l), and (Y_j, Y_k, Y_l, Y_p), respectively. This integral equation is also given in [19].

The efficacy functional of (2.27) is closely related to the generalized SNR functional of (2.7). Indeed, we can write

$$\eta(g) = \lim_{\theta \to 0} \frac{S_\zeta(g)}{\theta^2} \tag{2.30}$$

where $S_\zeta(g)$ is evaluated from (2.7) by using $f_1^{(1,2)}(x,y) = f^{(1,2)}(x-\theta, y-\theta)$, $f_0^{(1,2)}(x,y) = f^{(1,2)}(x,y)$, and similar relationships for the third- and fourth-order joint densities.

The efficacy has also a property similar to that of the SNR with the matched filter. If we multiply both sides of (2.28) by $g(x,y)$ and integrate, we find that, if g^* solves the integral equation, then (2.28) implies that

$$\mu_0'(g^*) = \sigma_0^2(g^*). \tag{2.31}$$

This gives the value

$$\eta(g^*) = \mu_0'(g^*) = -\iint g^*(x,y) \left[\frac{\partial}{\partial x} f^{(1,2)}(x,y) + \frac{\partial}{\partial y} f^{(1,2)}(x,y)\right] dx\, dy \tag{2.32}$$

of the efficacy evaluated at the optimal nonlinearity.

Under mismatch conditions, that is, when $g(x,y) \neq g^*(x,y)$, the efficacy $\eta(g)$ can be obtained from (2.10), where the numerator by

$$\mu_0'(g) = -\iint g(x,y) \left[\frac{\partial}{\partial x} f^{(1,2)}(x,y) + \frac{\partial}{\partial y} f^{(1,2)}(x,y)\right] dx\, dy \tag{2.33}$$

and the denominator is given by

$$\begin{aligned} \sigma_0^2(g) = &\iint g^2(x,y) f^{(1,2)}(x,y)\, dx\, dy \\ &+ \iiint g(x,y)\left[g(u,x) f^{(1;2,3)}(u,x,y) + g(y,u) f^{(1,2,3)}(x,y,u)\right] du\, dx\, dy \\ &+ \iiiint g(x,y) g(u,v) K(x,y,u,v)\, dx\, dy\, du\, dv. \end{aligned} \tag{2.34}$$

2.5 A Lower Bound for Worst-Case Performance

In our work on memoryless discriminators [53], we were able to obtain a nonlinearity which is robust in a minimax sense. We have not been able to extend these results fully to the one-step memory case. What we have obtained is a lower bound on performance, in terms of the SNR performance measure, which holds for every distribution in the uncertainty classes. This is not a minimax robustness result because we have not been able to demonstrate the existence of least favorable distributions whose optimal nonlinearity

achieves the lower bound. However, this information is useful, since the lower bound may be able to tell us that significant performance degradation from the nominal performance can be avoided by using the robust test. The treatment here is brief, since most of the work here is an extension of the results in [53].

The notation in this section is cumbersome. F_0 and F_1 denote the complete distributions under H_0 and H_1 (i.e. the complete set of finite dimensional distributions), while f_0 and f_1 denote the bivariate distributions of (X_1, X_2) under H_0 and H_1, respectively. We use $S_\zeta(g; F_0, F_1)$ to denote the performance measure (2.7) evaluated under the indicated distributions. A related performance measure $S'_\zeta(g; f_0, f_1)$, which involves only the bivariate distributions, is defined by

$$S'_\zeta(g; f_0, f_1) = \frac{|\mathrm{E}_1\, g(X_1, X_2) - \mathrm{E}_0\, g(X_1, X_2)|^2}{(1-\zeta)(1+2R_0)\mathrm{Var}_0\, g(X_1, X_2) + \zeta(1+2R_1)\mathrm{Var}_1\, g(X_1, X_2)}, \tag{2.35}$$

where $R_i = \sum_{j=2}^{m} r_{ij}$ are known constants. The parameters r_{ij} are defined below.

Uncertainty classes for the distributions are defined in the following way. The bivariate densities are assumed to belong to classes defined by ϵ-mixtures. (The results hold for other uncertainty classes, as well, such as the 2-alternating capacity classes [23]). Thus every bivariate density under H_i is assumed to have the form

$$f_i(x, y) = (1 - \epsilon_i) f_i^0(x, y) + \epsilon_i h_i(x, y) \tag{2.36}$$

where f_i^0 is a nominal density and h_i is an unknown density. The parameter $\epsilon_i \in (0, 1)$ determines the degree of uncertainty in the bivariate distribution. A further condition on the uncertainty classes, given below, completely defines them. This condition requires that the bound

$$\sup_g \frac{\mathrm{Cov}_i[g(X_1, X_2), g(X_{j+1}, X_{j+2})]}{\sqrt{\mathrm{Var}_i\, g(X_1, X_2)\mathrm{Var}_i\, g(X_{j+1}, X_{j+2})}} \leq r_{ij} \tag{2.37}$$

be satisfied for a known sequence r_{ij}, for $i = 0, 1$ and $j = 2, 3, \ldots, m$. The sequence of r parameters determines the degree of uncertainty in the higher-order distributions. Note that for ρ-mixing processes, we have $r_{ij} \leq \rho_j$ for all j.

For the performance measure $S_\zeta(g; f_0, f_1)$, which involves only the bivariate distributions, we may obtain a complete robustness result. One can show [53] that the least-

favorable bivariate densities $\hat{f}_0$, $\hat{f}_1$ are those given by Huber [22],

$$\begin{aligned} \hat{f}_0(x,y) &= \begin{cases} (1-\epsilon_0)f_0^0(x,y) & \text{if } f_1^0(x,y)/f_0^0(x,y) < c'' \\ (1/c'')(1-\epsilon_0)f_1^0(x,y) & \text{if } f_1^0(x,y)/f_0^0(x,y) \geq c'' \end{cases} \\ \hat{f}_1(x,y) &= \begin{cases} (1-\epsilon_1)f_1^0(x,y) & \text{if } f_1^0(x,y)/f_0^0(x,y) > c' \\ c'(1-\epsilon_1)f_0^0(x,y) & \text{if } f_1^0(x,y)/f_0^0(x,y) \leq c' \end{cases} \end{aligned} \tag{2.38}$$

where the constants c' and c'' are chosen so that the functions integrate to one. The nonlinearity $\hat{g}$ that is matched to these densities is minimax robust under the performance measure S'_ζ, that is, the inequalities

$$S'_\zeta(\hat{g}; f_0, f_1) \geq S'_\zeta(\hat{g}; \hat{f}_0, \hat{f}_1) \geq S'_\zeta(g; \hat{f}_0, \hat{f}_1) \tag{2.39}$$

are satisfied for any other densities f_0 and f_1 in the uncertainty class and for any other nonlinearity g. This nonlinearity is given by the solution of the integral equation

$$\hat{g}(x,y) = \frac{\hat{f}_1(x,y) - \hat{f}_0(x,y)}{A\hat{f}_1(x,y) + \hat{f}_0(x,y)} + \iint \frac{A\hat{f}_1(x,y)\hat{f}_1(u,v) + \hat{f}_0(x,y)\hat{f}_0(u,v)}{A\hat{f}_1(x,y) + \hat{f}_0(x,y)} \hat{g}(u,v)\, du\, dv \tag{2.40}$$

where

$$A = \frac{\zeta(1+2R_1)}{(1-\zeta)(1+2R_0)}. \tag{2.41}$$

The integral equation (2.40) may be obtained directly by the method used to obtain (2.20). The solution of this integral equation was obtained in [53] and is given by

$$\hat{g}(x,y) = \left[\iint \frac{\hat{f}_0(x,y)\hat{f}_1(x,y)}{A\hat{f}_1(x,y) + \hat{f}_0(x,y)}\, dx\, dy\right]^{-1} \frac{\hat{f}_1(x,y)}{A\hat{f}_1(x,y) + \hat{f}_0(x,y)} \tag{2.42}$$

We now obtain the lower bound for the complete uncertainty classes. The lower bound is given by

$$S_\zeta(\hat{g}; F_0, F_1) \geq S'_\zeta(\hat{g}; f_0, f_1) \geq S'_\zeta(\hat{g}; \hat{f}_0, \hat{f}_1). \tag{2.43}$$

The first inequality in (2.43) is obtained from the bounds (2.37) and the expression (2.6). The second inequality holds because of the minimax robustness of $\hat{g}$ over the ϵ-contamination classes for the bivariate densities, the proof of which is an extension of the proof given in [53] for univariate densities.

In order to demonstrate true minimax robustness for the complete uncertainty classes, one must construct least favorable multivariate distributions $\hat{F}_0$ and $\hat{F}_1$ that satisfy the bounds

$$S_\zeta(\hat{g}; F_0, F_1) \geq S_\zeta(\hat{g}; \hat{F}_0, \hat{F}_1) \geq S_\zeta(g; \hat{F}_0, \hat{F}_1) \tag{2.44}$$

for arbitrary distributions F_0, F_1, and an arbitrary nonlinearity g. The way to do this is to extend the least favorable bivariate densities $\hat{f}_0$ and f_1 to multivariate distributions such that equality holds in (2.37) for each j. If this this were the case, then we would have the equality

$$S'_\zeta(\hat{g}; \hat{f}_0, \hat{f}_1) = S_\zeta(\hat{g}; \hat{F}_0, \hat{F}_1). \tag{2.45}$$

Then, combining (2.40) and (2.43) would yield the left inequality in (2.39). We could also obtain the optimal nonlinearity $\hat{g}$ by the method of Section 2.3, and thus we would also have the right inequality. For memoryless nonlinearities, it is possible to construct such multivariate distributions [51] and thus prove minimax robustness [53]. It should be noted, however, that the distributions in the memoryless case possess a very peculiar property which is necessary for showing robustness inequalities. This property seems to be lost when dimensionality is extended.

2.6 Simulation Studies

2.6.1 Discrimination Between Rayleigh and Lognormal Correlated Time-Series

In Section 2.3, we provided a method for obtaining an approximate solution of the integral equation (2.20) by solving a system of linear equations (2.26). We used this method, with $N = 33$ nodes, to obtain (approximately) the optimal one-step memory nonlinearity for discrimination between a Rayleigh and lognormal process. We then evaluated the performance of the one-step memory discriminator through computer simulation. The results are presented here.

The discrimination problem we consider assumes a Rayleigh distribution under H_0 and a lognormal distribution under H_1. These two distributions are common in the modeling of radar signals. The Rayleigh process $\{X_k\}$ occurs as the envelope of a complex Gaussian signal, so that $X_k = \sqrt{Y_k^2 + Z_k^2}$ where $\{Y_k\}$ and $\{Z_k\}$ are the real and imaginary components of the underlying Gaussian processes. We assume that the distributions of Y_k and Z_k have mean $\mu_Y = \mu_Z = 0$, variance $\sigma_Y^2 = \sigma_Z^2 = 4$, and covariances $\mathrm{Cov}[Y_k, Y_{k+j}] = \mathrm{Cov}[Z_k, Z_{k+j}] = e^{-|j|/\tau_0}$, with $\tau_0 = 13$. The lognormal process $\{X_k\}$, which is sometimes used to model signals that have a have a heavy tail part in their distributions, is related to a Gaussian process $\{U_k\}$ through $X_k = \exp(U_k)$. We assume that $\mu_U = 0.8$, $\sigma_U^2 = 0.25$, and

$\text{Cov}[U_k, U_{k+j}] = e^{-|j|/\tau_1}$, with $\tau_1 = 130$. With these parameters, the Rayleigh and lognormal distributions have equal first and second moments. The tails are different, however, with the lognormal distribution having a heavier tail to the right (toward $+\infty$) and the Rayleigh distribution having a heavier tail to the left (toward zero). The correlation is considerably different, being very strong under H_1 (lognormal) and moderate under H_0 (Rayleigh). We assume that an m-dependent model with $m = 300$ is sufficient approximation for either hypothesis. The forms of the Rayleigh and lognormal second-, third-, and fourth-order densities are given in Appendix A.

Figure 2.1 shows the optimal one-step memory nonlinearity, which is symmetric in x_i and x_{k+1}. Figures 2.2 and 2.3 show plots of P_1 vs. P_0, where P_i is the probability of error when H_i is true, for three different discriminators. These discriminators are denoted $T_n^{(0)}$ for the discriminator that is optimal under the i.i.d. assumption (Eq. (2.3) with $g(x) = \log[f_1(x)/f_0(x)]$); $T_n^{(1)}$ for the memoryless discriminator (2.3) that is optimal under the performance measure (2.7) with $\zeta = \frac{1}{2}$; and $T_n^{(2)}$ for the optimal one-step memory discriminator obtained from the integral equation (2.20) with $\zeta = \frac{1}{2}$. The memoryless discriminator $T_n^{(1)}$ is obtained through the solution of an integral equation similar to (2.20), as shown in [53]. Figure 2.2 was generated from 10^6 simulations with $n = 25$ samples for the test statistic. The superiority of $T_{25}^{(2)}$ is evident. This superiority is even more evident in Figure 2.3, also generated from 10^6 simulations but with $n = 50$ samples.

Generally, 10^6 trials is not sufficient to accurately approximate error probabilities whose magnitude is on the order of 10^{-4} unless importance sampling is used. Since we did not use importance sampling, the curve for $T_n^{(2)}$ in Figure 2.3 should not be regarded as an accurate estimation of the true error probability curve. Nevertheless, the plot does indicate a substantial improvement in performance obtained by the one-step memory discriminator over the memoryless discriminator.

2.6.2 Detection of a Weak-Signal in Correlated Cauchy Noise

The univariate density for Cauchy noise is $f(x) = [\pi(1+x^2)]^{-1}$. This distribution has extremely heavy tails and is a good model for impulsive noise. Correlated Cauchy noise may be obtained from a memoryless transformation of a Gaussian process. For our model, $\{Z_k\}$ is assumed Gaussian with zero mean, unit variance, and covariance $\text{Cov}[Z_k, Z_{k+j}] = e^{-|j|/\tau}$, and the Cauchy noise process is obtained by $Y_k = \tan\left[\pi \,\text{erf}(Z_k/\sqrt{2})/2\right]$. The forms

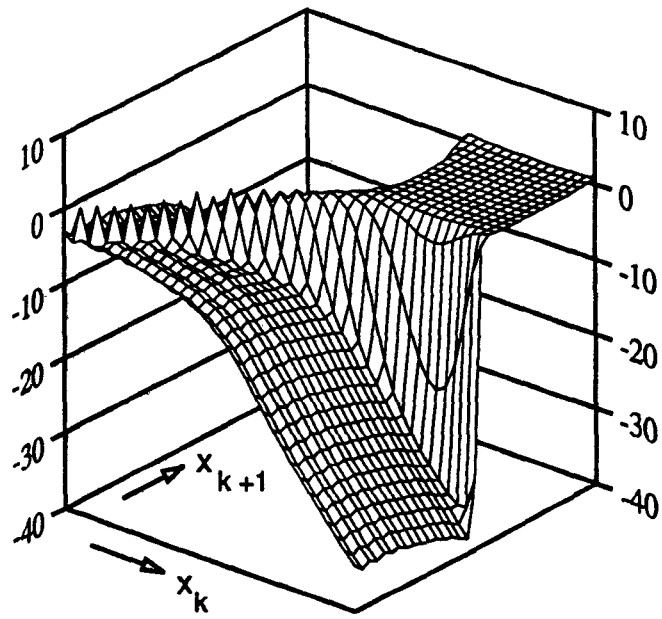


Figure 2.1: Graph over the region $[0, 16] \times [0, 16]$ of the optimal one-step memory nonlinearity for discrimination of Rayleigh vs. lognormal processes.

of the Cauchy second-, third-, and fourth-order densities are given in Appendix B.

Figures 2.4 and 2.5 show graphs of the nonlinearities for $\tau = 1$ and $\tau = 15$. Table 2.1 shows the value of the efficacy for several detectors and values of τ. The values are normalized in Table 2.2 to show the ARE to the locally optimal i.i.d. (memoryless) detector with $g(x) = -f'(x)/f(x)$. The Markov detector is a one-step memory detector with

τ	i.i.d. L.O.	optimal memoryless	Markov L.O.	optimal one-step
1	0.3320	0.3733	0.4003	0.4242
5	0.0922	0.1512	0.3921	0.4253
15	0.0316	0.0567	0.4138	0.4260
30	0.0158	0.0288	0.4225	0.4270

Table 2.1: Efficacies of various detectors vs. correlation strength τ.

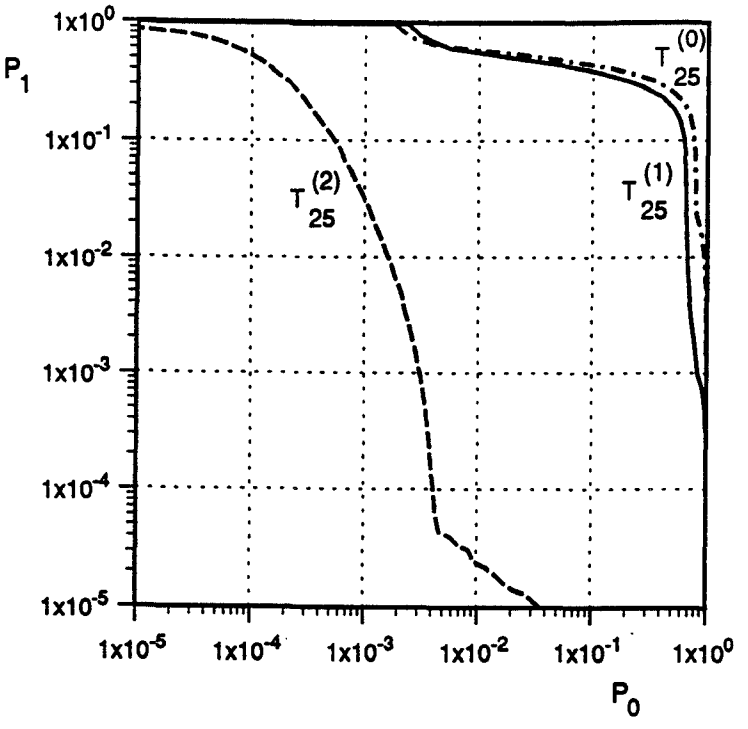


Figure 2.2: Plots of the error probabilities for the i.i.d. discriminator $T^{(0)}$, the optimal memoryless discriminator $T^{(1)}$, and the optimal one-step memory discriminator $T^{(2)}$ for a sample size of $n = 25$.

τ	i.i.d. L.O.	optimal memoryless	Markov L.O.	optimal one-step
1	1.000	1.124	1.206	1.278
5	1.000	1.640	4.253	4.614
15	1.000	1.794	13.10	13.48
30	1.000	1.817	26.67	26.95

Table 2.2: ARE of various detectors relative to the i.i.d. locally optimal detector vs. correlation strength τ.

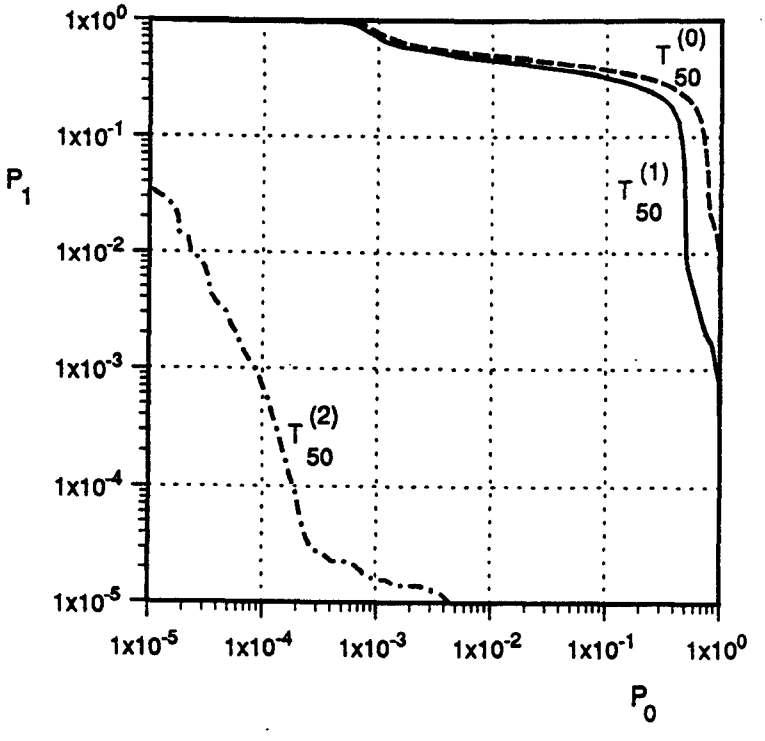


Figure 2.3: Plots of the error probabilities for the i.i.d. discriminator $T^{(0)}$, the optimal memoryless discriminator $T^{(1)}$, and the optimal one-step memory discriminator $T^{(2)}$ for a sample size of $n = 50$.

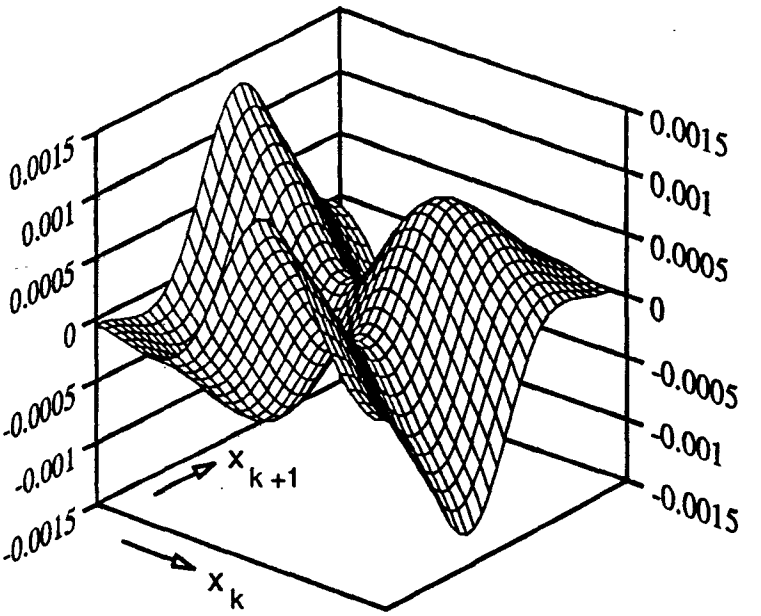


Figure 2.4: Graph over the region $[-233, 233] \times [-233, 233]$ of the optimal one-step memory nonlinearity for signal detection in Cauchy noise with $\tau = 1$.

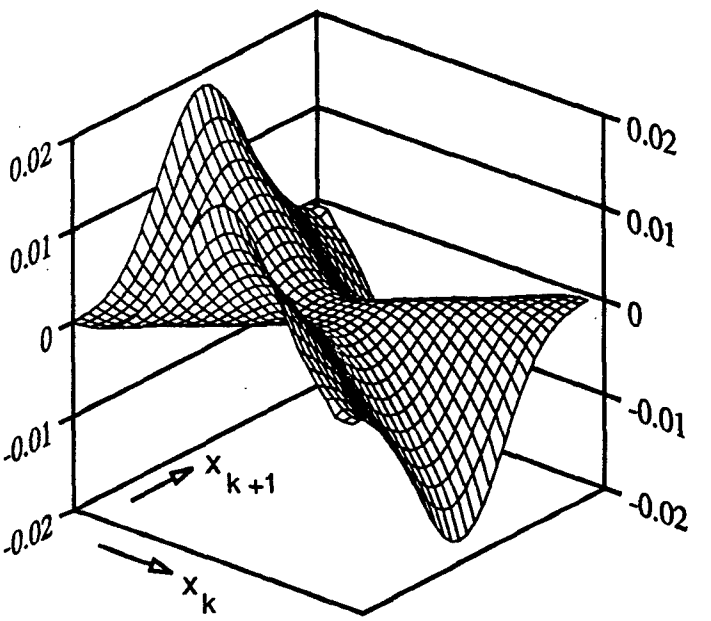


Figure 2.5: Graph over the region $[-233, 233] \times [-233, 233]$ of the optimal one-step memory nonlinearity for signal detection in Cauchy noise with $\tau = 15$.

$$\begin{aligned} g(x,y) &= \frac{d}{d\theta}\log\left[\frac{f(x-\theta,y-\theta)f(x)}{f(x,y)f(x-\theta)}\right]_{\theta=0} \\ &= -\frac{\frac{\partial}{\partial x}f(x,y)+\frac{\partial}{\partial y}f(x,y)}{f(x,y)}+\frac{f'(x)}{f(x)}, \end{aligned} \tag{2.46}$$

which is the locally optimal detector if the noise process is Markov. It is interesting to note that the efficacy of the memoryless detectors decreases with increasing τ, while the efficacy of the one-step memory detectors increases. Since the Markov detector behaves like a memoryless detector for relatively weak correlation and a one-step memory detector for stronger correlation, its efficacy at first decreases and then increases, as τ increases.

2.7 Discussion

The contribution of our research has been to extend results obtained for memoryless discriminators to discriminators involving one-step memory. In Section 2.3, we showed that the optimal one-step memory nonlinearity can be obtained as the solution of the linear integral equation (2.20). We also showed how an approximate solution of the integral equation can be obtained by solving a system of linear equations. This numerical approximation actually involves a quantization of the nonlinearity, which we do not claim to be optimal. However, our simulation results demonstrate a significant improvement over memoryless discriminators even for a crude solution with $N = 33$ nodes. Evidently, any loss of accuracy in the approximation was not significant enough to override the added discrimination power contributed by the memory. This numerical approximation to the one-step memory nonlinearity involved the solution of a 1089×1089 system of linear equations, which is well within the computation capability of many computer systems. In our preliminary results, we used only 17 nodes for the integration and still obtained a significant improvement over the optimal memoryless discriminator. We actually used interpolation to generate values of g, though, and not quantization, in our simulations. Results on optimal quantization may be possible, although they are more complex than results for quantization of the memoryless nonlinearities.

We have also been able to apply our results to the problem of weak-signal detection. We showed that the optimal one-step memory nonlinearity for detection under the ARE criterion is given by the solution of the linear integral equation (2.28). The performance of the optimal one-step memory detector as compared to the optimal memoryless detector

under the ARE performance measure is significantly better and improves with increasing correlation.

Results for robustness for memoryless discriminators do not completely extend to the case of memory, as we showed in Section 2.5. What we obtained there is a one-step memory nonlinearity $\hat{g}$ and a lower bound on its performance when the distributions are known to belong to given uncertainty classes.

Our simulation results demonstrate that the optimal one-step memory discriminator performs significantly better than the optimal memoryless discriminator. Similarly, in the weak-signal detection problem, our numerical results show that the ARE of the optimal one-step memory detector is significantly better than that of the optimal memoryless detector.

Chapter 3

Models for Positive-Valued Random Processes

3.1 Introduction

The generalized SNR utilized in Chapter 2 is an asymptotic performance measure that depends on the asymptotic mean and variance of the test statistic. Since asymptotic results normally apply only to problems involving large sample sizes, the usefulness of the performance measure for small or even moderate sample sizes is questionable. This fact provides the primary motivation for the approach taken in this chapter and the one following, where we consider discrimination methods that are applicable to small and [redacted] sample sizes.

It is well known that the likelihood ratio test (LRT) is optimal with respect to the error probabilities of a discriminator. In general, we prefer a performance criterion based on the error probabilities, such as the Bayes or Neyman-Pearson criterion, to a generalized SNR performance criterion, since the error probabilities are more physically meaningful. Therefore, it seems logical to prefer a LRT discriminator whenever possible. There are, however, valid reasons for preferring a discriminator other than the LRT and a performance criterion other than the error probabilities. One such reason is that the LRT may be too difficult to implement because of the difficulty in computing the multivariate probability density functions (pdfs) of the processes. While there are other valid reasons for not using the LRT, it is this one reason in particular—the unavailability of easily computable

multivariate pdfs—that we are most concerned with here.

In both this chapter and the following one, we derive multivariate pdfs that are suitable for use in an LRT discriminator. Assuming the signals involved have such pdfs, the LRT discriminator is optimal under any criterion directly related to the error probabilities, including the Bayes, Neyman-Pearson, and minimax criteria. Two issues that we must consider here include the computability and the accuracy of the pdfs. As we pointed out above, computability is essential to the implementation of the LRT discriminator. Some radar applications require that discrimination be performed in real time on minimal hardware. On the other hand, the performance of the LRT discriminator depends on how accurately the assumed pdfs match the true pdfs of the signals. Because the accuracy of a pdf is difficult to assess, we partially side step the accuracy issue by considering instead the flexibility of the pdfs. Our rationale here is that the more flexible a pdf is, the more likely it is that one will obtain a good fit to observed data (that is, the more accurate the pdf will be). As a starting point, we make the reasonable assumption that a minimally flexible multivariate pdf should be capable of matching specified univariate marginal pdfs and should provide a fairly flexible dependency structure. We must be careful, however, not to put to much emphasis on the accuracy or flexibility issue at the expense of the computability issue. We may tend to think, somewhat informally perhaps, in terms of an accuracy/computability trade-off.

The problem of constructing a multivariate pdf with specified univariate pdfs, which is essentially the problem considered here, has been previously studied from two different perspectives. First, there are the statisticians, who are concerned with multivariate distributions that can explain the statistical relationships between several different sets of measurements. The multivariate pdfs that have been devised for such applications are typically bivariate pdfs or pdfs of some other low dimensionality, although some are capable of being extended to arbitrary dimensionality. Kimeldorf and Sampson [29] provide a brief discussion of some available bivariate uniform distributions. Johnson [24] is a good reference for multivariate distributions that are suitable for Monte Carlo studies. Distributions of low dimensionality, such as bivariate distributions, may be used to synthesize stationary distributions of arbitrary dimension by forming products of conditional pdfs. For example, we may form the conditional pdf $f(x|y) = f(x,y)/f(y)$ whenever a bivariate form of a given univariate pdf is available. Apart from the statisticians, there are researchers who are primarily concerned with time series analysis. Since the applications of time series analysis

often include such areas as spectral estimation and forecasting, the univariate distribution is not always a primary concern. Often, univariate distributions are assumed to be Gaussian; on other occasions, a memoryless transformation is applied to the observed data and the transformed data is assumed to be Gaussian. A notable exception to this approach in time series modeling is the work of Lawrance and Lewis [32], who devised an autoregressive time series model with exponential innovations and exponential marginal pdf for the model output. Their work has been praised for its intentions but not otherwise well received for various practical reasons. (See the discussion that immediately follows the paper in the journal.)

In signal processing applications, multivariate distributions are often needed for modeling or for simulations. In both cases, whenever multivariate distributions with specified univariate pdfs are needed, transformed Gaussian distributions are used almost exclusively. The transformed Gaussian does indeed provide a flexible multivariate pdf, since it can match any specified univariate marginals and any specified (legitimate) autocovariance function [38], [34]. Considering the flexibility of this method and the tractability it affords, it is not surprising that alternatives are uncommon. While one could use another underlying distribution instead of a Gaussian and still use the transformation to shape the marginals, the problem lies in obtaining a suitable alternative to the multivariate Gaussian distribution. There really are no true multivariate distributions that provide the flexibility and tractability of the multivariate Gaussian. Cantrell [6] uses a bivariate Cauchy distribution where the correlated Cauchy variables are obtained as a moving average of independent Cauchy variables. Cantrell's method is easily extended to arbitrary dimensionality; however, the distribution is limited in its ability to match an arbitrary autocovariance function.

In this chapter, we propose that the multivariate χ^2 distribution be considered as an alternative to the multivariate Gaussian as an underlying distribution for a transformation model. The χ^2 has several advantages. First, it is easy to generate data from the distribution that can be used in Monte Carlo studies. Second, the multivariate distributions are able to match any autocovariance function, as in the transformed Gaussian case. And third, the multivariate χ^2 with $d = 2$ degrees of freedom combined with a simple power transformation yields a multivariate Weibull distribution, which is commonly used in radar problems.

By assuming that the underlying Gaussian processes from which the χ^2 is derived are autoregressive, we are able to simplify considerably the computation of the multivariate

pdf. This is perhaps the major result of this chapter. We begin the chapter with a discussion of the preliminary theory upon which the models we propose are based. Next, we discuss methods for computing the multivariate χ^2 pdf, as is required for an LRT discriminator. We look at some examples to assess just how accurate the approximate computational methods are. We then discuss a method for estimating the parameters of the pdf from sample moments, a method adapted from a least squares method used to estimate parameters of an ARMA model. After this, we look at some contour plots that are helpful in showing the differences in the available models. Finally, we put the theoretical ideas into practice by actually implementing LRT discriminators and testing them through simulations. The simulations demonstrate the usefulness of the methods as well as call our attention to some important practical matters.

3.2 Preliminary Theory

In this chapter, three different ideas are combined to produce a flexible and useful model. First, the idea of a memoryless, invertible transformation is used to provide flexibility in choosing the univariate marginal pdf of the model output process. Second, the idea of a χ^2 distribution is introduced to provide flexibility in modeling the dependency structure of the output variables. Finally, the idea of a stationary, autoregressive process is introduced to permit efficient computation of the multivariate pdf. We now discuss these three ideas in detail.

3.2.1 Transformation Models

While there are many univariate distributions that are commonly used in statistics, few of these are used in multivariate forms. Compared to the large arsenal of univariate distributions, only a handful of true multivariate distributions are commonly recognized, such as the Gaussian, Wishart, etc., and out of these, only the Gaussian has a relatively simple multivariate probability density function (pdf). Nevertheless, a straightforward method for obtaining a multivariate pdf having specified univariate marginal pdfs is to use a memoryless, invertible transformation in conjunction with a known multivariate pdf. In this context, we may consider the multivariate distribution obtained by the transformation of a Gaussian process to be the prototype. Other transformation models may be obtained by substituting another multivariate distribution for the Gaussian.

Let $\{X_t\}$ be the observed stationary discrete-time random process, with known univariate pdf f_x and cumulative distribution function (cdf) F_x, for which we desire to obtain a multivariate pdf. Let $\{Z_t\}$ denote a random process that has a known multivariate pdf, such as a multivariate Gaussian. Using the memoryless transformation approach, we define a one-to-one (invertible) memoryless nonlinearity ψ such that $X_t = \psi^{-1}(Z_t)$ has the cdf F_x. Letting $f_z^{(n)}$, and $f_x^{(n)}$ denote, respectively, the n-dimensional pdfs for $\{Z_t\}$ and $\{X_t\}$, we have

$$f_x^{(n)}(\mathbf{x}) = f_z^{(n)}(\mathbf{z}) \prod_{t=1}^{n} \psi'(x_t) \tag{3.1}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{z} = [\psi(x_1), \ldots, \psi(x_n)]$. If the cdfs F_x and F_z are invertible, then one can always obtain such an invertible transformation ψ by setting

$$\psi(x) = F_z^{-1}\left[F_x(x)\right]$$

As a first attempt at defining a multivariate pdf, the pdf obtained by a transformed Gaussian is surprisingly simple and flexible. Using it, one can accurately model not only the univariate distributions, but also the covariance structure as well. Given the $n \times n$ covariance matrix of the observed process, one can obtain a multivariate distribution via a transformed Gaussian that matches it. In this chapter, we regard the transformed Gaussian model as our prototype, and we move on to explore more sophisticated models that can potentially offer a better fit to the true distribution of the observed process.

3.2.2 The χ^2 Distribution

The assumption of a multivariate normal distribution for $\{Z_t\}$ is most convenient in that the resulting multivariate pdf $f_x^{(n)}$ for the observations takes a relatively simple form. This also permits a relatively simple implementation of the LRT discriminator. From the standpoint of simplicity or convenience, the multivariate distribution obtained as a transformation of a multivariate Gaussian is hard to beat. We ought to consider, though, the possibility of obtaining a better fit to the observations by assuming an underlying distribution other than the Gaussian. If the observation process $\{X_t\}$ takes only positive values, then it is not hard to see why we might obtain a better fit by assuming another distribution. Since the Gaussian pdf is symmetric and positive over the entire real line, transforming Gaussian random variables into random variables that take only positive values results in

a substantially distorted dependency structure. By contrast, transforming, say, exponentially distributed random variables into positive-valued random variables results in a lesser distortion of the dependency structure. These two possible transformation models stand in stark contrast to each other. On the one hand, we transform a two-sided process into a one-sided process, while on the other hand we transform a one-sided process into another one-sided process. That the dependency structure would be significantly different in these two cases seems intuitive. If we take the underlying distribution to be a χ^2 distribution, as we do in this chapter, then we can again match the univariate marginals and autocovariance function of the observation process as in the Gaussian case. Thus any difference in the dependency structure, as compared to the Gaussian, may be observed only by going beyond the covariance values.

The χ^2 distribution, while being considerably more complex than the Gaussian, nevertheless does provide some degree of tractability. In particular, it is possible to match any autocovariance function, as in the case of the transformed Gaussian. Furthermore, it is relatively simple to generate samples from a multivariate χ^2 distribution. This property is important if the distribution is needed for generating random variables for Monte Carlo simulations. It is also worth mentioning that the Weibull distribution, which is commonly used in radar modeling, may be obtained as a simple power transformation of the χ^2 distribution with two degrees of freedom.

The χ^2 distribution arises frequently in statistics as the sum of squares of independent Gaussian random variables with zero means. Let $\{\mathbf{U}_t, 1 \leq t \leq n\}$ be a collection of d-dimensional random vectors $\mathbf{U}_t = (U_{t1}, \ldots, U_{td})$, where each U_{ti} has a standard Gaussian distribution with zero mean and unit variance. The sequence $\{\mathbf{U}_t, 1 \leq t \leq n\}$ is best thought of as a collection of d independent Gaussian sequences $\{U_{ti}, 1 \leq t \leq n\}$, $1 \leq i \leq d$, that are identically distributed. We denote the common covariance matrix of each of these scalar sequences by $\mathbf{Q}^{-1}$. (We are primarily concerned with the inverse $\mathbf{Q}$ of the covariance matrix.) Now, the random variable

$$Z_t = \sum_{i=1}^{d} U_{ti}^2, \tag{3.2}$$

is a sum of independent random variables. By definition, Z_t has a univariate χ^2 distribution with d degrees of freedom, denoted $\chi^2(d)$. The sequence $\{Z_t, 1 \leq t \leq n\}$, with Z_t defined by (3.2), has a multivariate $\chi^2(d)$ distribution. It is worth pointing out that the $\chi^2(2)$

distribution is the well-known exponential distribution.

The noncentral χ^2 distribution arises in a similar way as the sum of squares of Gaussian variables with non-zero means. Thus, by definition, Z_t has a noncentral χ^2 distribution with d degrees of freedom and noncentrality parameter α^2, denoted $\chi'^2(d, \alpha^2)$, when

$$Z_t = \sum_{i=1}^{d} (U_{ti} + \mu_i)^2, \tag{3.3}$$

and

$$\alpha^2 = \sum_{i=1}^{d} \mu_i^2. \tag{3.4}$$

The representation (3.3) is not unique, since the distribution depends only on $\mu_1, \ldots, \mu_d$ via the sum of their squares. Thus $\{Z_t, 1 \le t \le n\}$ has the same distribution for any $\mu_1, \ldots, \mu_d$ that satisfy (3.4). We focus primarily on the central χ^2 distribution in this chapter, since it is simpler and suits our purposes well. Most of the results, such as the methods described for computing the multivariate pdf, may be adapted to the noncentral χ^2, and there certainly are other applications where the ability to compute the multivariate noncentral χ^2 pdf is invaluable. This topic is discussed further in the final section of this chapter.

To simplify our derivation of the multivariate pdf, let us assume for now that $d = 2$. We will work through this derivation and then extend the result to cover other values for d. Therefore, let $Z_t = U_t^2 + V_t^2$ where $\{U_t\}$ and $\{V_t\}$ are identically distributed sequences of zero-mean Gaussian random variables that are independent of each other (that is, $\{U_t\}$ is independent of $\{V_t\}$). The sequence $\{Z_t\}$ has a multivariate exponential pdf that we now derive. Let $\mathbf{Q}$ be the inverse of the common covariance matrix for both $\mathbf{U} = (U_1, \ldots, U_n)$ and $\mathbf{V} = (V_1, \ldots, V_n)$, let $W_t = \sqrt{Z_t}$, let $E = \{\mathbf{u}, \mathbf{v} : \sqrt{u_t^2 + v_t^2} \le w_t\}$, and let $r_t \cos\theta_t = u_t$, $r_t \sin\theta_t = v_t$. To derive the pdf for $\mathbf{W} = (W_1 \ldots, W_n)$, we have

$$P\{W_t \le w_t, 1 \le t \le n\} = P\{\sqrt{U_t^2 + V_t^2} \le w_t, 1 \le t \le n\} \tag{3.5a}$$

$$= \int_E f_u(\mathbf{u}) f_v(\mathbf{v})\, d\mathbf{u}\, d\mathbf{v} \tag{3.5b}$$

$$= (2\pi)^{-n} |\mathbf{Q}| \int_E \exp\left\{-\tfrac{1}{2}\mathbf{u}^T\mathbf{Q}\mathbf{u} - \tfrac{1}{2}\mathbf{v}^T\mathbf{Q}\mathbf{v}\right\} d\mathbf{u}\, d\mathbf{v} \tag{3.5c}$$

$$= (2\pi)^{-n} |\mathbf{Q}| \int_E \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(u_i u_j + v_i v_j) q_{ij}\right\} d\mathbf{u}\, d\mathbf{v} \tag{3.5d}$$

$$= (2\pi)^{-n} |\mathbf{Q}| \int_0^{w_1} \cdots \int_0^{w_n} \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} r_i r_j \cos(\theta_i - \theta_j) q_{ij}\right\}$$

$$\times\, d\theta_1 \cdots d\theta_n \left(\prod_{i=1}^{n} r_i \right) dr_1 \cdots dr_n. \tag{3.5e}$$

From (3.5e) we can see that the pdf for $\mathbf{W}$ is given by

$$f_w^{(n)}(\mathbf{w}) = (2\pi)^{-n}|\mathbf{Q}| \left(\prod_{i=1}^{n} w_i \right) \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\left\{ -\tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \cos(\theta_i - \theta_j) q_{ij} \right\} d\theta_1 \cdots d\theta_n.$$

Therefore, the pdf for $\mathbf{Z} = (Z_1, \ldots, Z_n)$ is

$$f_z^{(n)}(\mathbf{z}) = (4\pi)^{-n}|\mathbf{Q}| \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\left\{ -\tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{z_i z_j} \cos(\theta_i - \theta_j) q_{ij} \right\} d\theta_1 \cdots d\theta_n. \tag{3.6}$$

Of particular interest to us, however, is the pdf for $\mathbf{Z}$ given in the form

$$f_z^{(n)}(\mathbf{z}) = (4\pi)^{-n}|\mathbf{Q}| \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\left\{ -\tfrac{1}{2}\mathbf{u}^T\mathbf{Q}\mathbf{u} - \tfrac{1}{2}\mathbf{v}^T\mathbf{Q}\mathbf{v} \right\} d\theta_1 \cdots d\theta_n \tag{3.7}$$

where

$$\begin{aligned} u_t &= \sqrt{z_t} \cos\theta_t, \\ v_t &= \sqrt{z_t} \sin\theta_t. \end{aligned} \tag{3.8}$$

This form is obtained by backtracking a bit in the derivation (3.5).

Straightforward modifications may be made to derive the pdf for the noncentral χ^2 distribution with two degrees for freedom. In (3.5d) we replace u_t by $u_t - \mu_u$ and v_t by $v_t - \mu_v$. After the change to polar coordinates, it is also useful to make the substitutions $\mu_u = \alpha\cos\omega$ and $\mu_v = \alpha\sin\omega$. In the final result, we have the following expression for the pdf for $\chi'^2(2, \alpha^2)$.

$$\begin{aligned} f_z(\mathbf{z}) = (4\pi)^{-n}|\mathbf{Q}| \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\Bigg\{ -\tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} \Big[\sqrt{z_i z_j} \cos(\theta_i - \theta_j) \\ - \alpha\sqrt{z_i}\cos(\omega - \theta_i) - \alpha\sqrt{z_j}\cos(\omega - \theta_j) + \alpha^2 \Big] \Bigg\} d\theta_1 \cdots d\theta_n. \end{aligned} \tag{3.9}$$

Note that the value of the pdf depends on α but not on ω. Thus we may simplify the expression further by setting $\omega = 0$, which yields

$$\begin{aligned} f_z(\mathbf{z}) = (4\pi)^{-n}|\mathbf{Q}| \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\Bigg\{ -\tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} \Big[\sqrt{z_i z_j} \cos(\theta_i - \theta_j) \\ - \alpha\sqrt{z_i}\cos(\theta_i) - \alpha\sqrt{z_j}\cos(\theta_j) + \alpha^2 \Big] \Bigg\} d\theta_1 \cdots d\theta_n. \end{aligned} \tag{3.10}$$

The pdf may also be written in the form (3.7), where

$$\begin{aligned} u_t &= \sqrt{z_t}\cos\theta_t - \mu_u, \\ v_t &= \sqrt{z_t}\sin\theta_t - \mu_v. \end{aligned} \tag{3.11}$$

The derivation for $d > 2$ is similar to the case of $d = 2$, only more complicated. Analogous to the change to polar coordinates in (3.5e), at some point in the derivation a change is made to spherical coordinates for $d = 3$ or to generalized spherical coordinates for $d > 3$. This method is described in detail in [41]. Generalized spherical coordinates [54] are not unique, but any representation will suit our purposes. A form given in [41] is

$$\begin{aligned} x_j &= r\left(\prod_{k=1}^{j-1}\sin\phi_k\right)\cos\phi_j, \qquad 1 \le j \le d-2 \\ x_{d-1} &= r\left(\prod_{k=1}^{d-2}\sin\phi_k\right)\cos\theta \\ x_d &= r\left(\prod_{k=1}^{d-2}\sin\phi_k\right)\sin\theta \end{aligned} \tag{3.12}$$

where $r^2 = x_1^2 + \cdots + x_d^2 \ge 0$, $0 \le \phi_k < \pi$, and $0 \le \theta < 2\pi$. The Jacobian of the transformation is

$$J(r, \theta, \phi_1, \ldots, \phi_{d-2}) = r^{d-1}\prod_{k=1}^{d-2}\sin^{d-1-k}\phi_k. \tag{3.13}$$

Though the expression for the multivariate pdf for $d > 3$ is complex, the idea is fairly simple and is clearly illustrated in the following theorem, in which the integration is expressed as an expectation of a function of random variables that are uniformly distributed on the d-dimensional unit hypersphere.

Theorem 3.1 *Let $\{\mathbf{Y}_t, 1 \le t \le n\}$ be a sequence of independent and identically distributed random vectors from $\mathbb{R}^d$ such that $\mathbf{Y}_t = (Y_{t1}, \ldots, Y_{td})$ is uniformly distributed over the d-dimensional unit sphere $S = \{\mathbf{y} \in \mathbb{R}^d : y_1^2 + \cdots + y_d^2 = 1\}$. Then the n-dimensional pdf for the $\chi'^2(d, \alpha^2)$ distribution may be written*

$$f_z(\mathbf{z}) = \frac{|\mathbf{Q}|^{d/2}}{2^{nd/2}\Gamma(d/2)^n}\left(\prod_{t=1}^{n} z_t^{(d-2)/2}\right)\mathrm{E}\left[\exp\left\{-\tfrac{1}{2}\sum_{k=1}^{d}\mathbf{U}_k^T\mathbf{Q}\mathbf{U}_k\right\}\right] \tag{3.14}$$

where

$$U_{kt} = Y_{tk}\sqrt{z_t} - \mu_k. \tag{3.15}$$

and $\alpha^2 = \sum_{k=1}^{d}\mu_k^2$.

Proof. Let $\Phi_1, \Phi_2, \ldots, \Phi_{d-2}, \Theta$ be (independent) random variables having the joint pdf

$$f(\phi_1, \ldots, \phi_{d-2}, \theta) = \frac{\Gamma(d/2)}{2\pi^{d/2}} \sin^{d-2}\phi_1 \sin^{d-3}\phi_2 \cdots \sin\phi_{d-2},$$

$$0 \le \phi_k < \pi, \quad 0 \le \theta < 2\pi.$$

Then, as shown in [24, p. 126], the coordinates

$$\begin{aligned} Y_1 &= \cos\Phi_1 \\ Y_2 &= \sin\Phi_1 \cos\Phi_2 \\ &\vdots \\ Y_{d-1} &= \sin\Phi_1 \cdots \sin\Phi_{d-2}\cos\Theta \\ Y_d &= \sin\Phi_1 \cdots \sin\Phi_{d-2}\sin\Theta \end{aligned}$$

are uniformly distributed on the surface the d-dimensional unit hypersphere.

Now let $\mathbf{V}_1, \ldots, \mathbf{V}_d$ be Gaussian vectors such that $Z_t = V_{1t}^2 + \cdots + V_{dt}^2$ has the desired $\chi'^2(d, \alpha^2)$ distribution. Let $\mathbf{U}_k = \mathbf{V}_k - \mu_k$, $1 \le k \le d$, be the centered Gaussian random variables obtained by subtracting the mean. Let $W_t = \sqrt{Z_t}$ and $E = \{W_t \le w_t, 1 \le t \le n\}$. Then, proceeding as in (3.5), we have the following result, where we introduce the generalized spherical coordinates (3.12) for $v_{k,t}$:

$$\begin{aligned} &P\{W_t \le w_t, 1 \le t \le n\} \\ &\quad = (2\pi)^{-nd/2}|\mathbf{Q}|^{d/2} \int_E \exp\left\{-\tfrac{1}{2}\sum_{k=1}^{d}(\mathbf{v}_k - \mu_k)^T\mathbf{Q}(\mathbf{v}_k - \mu_k)\right\} d\mathbf{v}_1 \cdots d\mathbf{v}_d \qquad (3.16a) \\ &\quad = (2\pi)^{-nd/2}|\mathbf{Q}|^{d/2} \int_0^{w_n} r_n^{d-1}\, dr_n \cdots \int_0^{w_1} r_1^{d-1}\, dr_1 \qquad (3.16b) \\ &\qquad \times \int_{E_n} J'(\mathbf{s}_n)\, d\mathbf{s}_n \cdots \int_{E_1} J'(\mathbf{s}_1)\, d\mathbf{s}_1 \exp\left\{-\tfrac{1}{2}\sum_{k=1}^{d}(\mathbf{v}_k - \mu_k)^T\mathbf{Q}(\mathbf{v}_k - \mu_k)\right\} \end{aligned}$$

where

$$\begin{aligned} \mathbf{s}_t &= (\phi_{1t}, \ldots, \phi_{d-2,t}, \theta_t), \\ J'(\mathbf{s}_t) &= \prod_{k=1}^{d-2} \sin^{d-1-k}\phi_{kt}, \\ E_t &= \{0 \le \phi_{1t} < \pi, \ \ldots, \ 0 \le \phi_{d-2,t} < \pi, \ 0 \le \theta_t < 2\pi\}. \end{aligned}$$

In (3.16b), v_{kt} is understood to be a function of r_t and $\mathbf{s}_t$. Then, the pdf for $\mathbf{W}$ is

$$f_w(\mathbf{w}) = (2\pi)^{-nd/2}|\mathbf{Q}|^{d/2}\prod_{t=1}^{n} w_t^{d-1}$$
$$\times \int_{E_n} J'(\mathbf{s}_n)\,d\mathbf{s}_n \cdots \int_{E_1} J'(\mathbf{s}_1)\,d\mathbf{s}_1 \exp\left\{-\tfrac{1}{2}\sum_{k=1}^{d}(\mathbf{v}_k-\mu_k)^T\mathbf{Q}(\mathbf{v}_k-\mu_k)\right\}$$

and the pdf for $\mathbf{Z}$ is

$$f_z(\mathbf{z}) = (2\pi)^{-nd/2}|\mathbf{Q}|^{d/2}\prod_{t=1}^{n} z_t^{(d-1)/2}\prod_{t=1}^{n}\tfrac{1}{2}z_t^{-1/2}$$
$$\times \int_{E_n} J'(\mathbf{s}_n)\,d\mathbf{s}_n \cdots \int_{E_1} J'(\mathbf{s}_1)\,d\mathbf{s}_1 \exp\left\{-\tfrac{1}{2}\sum_{k=1}^{d}(\mathbf{v}_k-\mu_k)^T\mathbf{Q}(\mathbf{v}_k-\mu_k)\right\}$$
$$= \frac{|\mathbf{Q}|^{d/2}}{2^{nd/2}\Gamma(d/2)^n}\prod_{t=1}^{n} z_t^{(d-2)/2}\left(\frac{\Gamma(d/2)}{2\pi^{d/2}}\right)^n$$
$$\times \int_{E_n} J'(\mathbf{s}_n)\,d\mathbf{s}_n \cdots \int_{E_1} J'(\mathbf{s}_1)\,d\mathbf{s}_1 \exp\left\{-\tfrac{1}{2}\sum_{k=1}^{d}\mathbf{u}_k^T\mathbf{Q}\mathbf{u}_k\right\}$$

where $\mathbf{u}_k = \mathbf{v}_k - \mu_k$. Now, each integral of the form

$$\frac{\Gamma(d/2)}{2\pi^{d/2}}\int_{E_t}[\,\cdot\,]J'(\mathbf{s}_t)\,d\mathbf{s}_t$$

may be written as an expectation involving the variables $\Phi_{1t},\ldots,\Phi_{d-2,t},\Theta_t$. By a change of variables, we may write the expectation in terms of the rectangular coordinate random variables $\mathbf{Y}_t$. Thus we achieve the desired result. Q.E.D.

The computation of the multivariate $\chi^2(d)$ pdf requires that one evaluate a rather difficult multidimensional integral. The difficulty in evaluating this integral increases significantly with increasing d. For this reason, it is important to note that the $\chi^2(d)$ tends to a multivariate Gaussian distribution as $d \to \infty$. This fact follows from the multivariate form of the central limit theorem. Generally, convergence in the central limit theorem is more rapid for symmetric distributions and slower for distributions that are skewed. Since the $\chi^2(d)$ distribution arises as a sum of squared Gaussian random variables, which are clearly positively skewed, we expect the convergence to Gaussianity under the central limit theorem to be slow. Nevertheless, we focus our attention on the case of $d = 2$ and $d = 3$, not only for the sake of making the computation manageable, but also since these distributions provide the greatest contrast to the multivariate Gaussian distribution.

3.2.3 Autoregressive Models

We require the assumption of an autoregressive (AR) model in order to streamline the computation of the multivariate χ^2 pdfs. Without such an assumption, the complexity involved in computing the n-dimensional pdf typically increases exponentially in n. We shall see, however, that the complexity is linear in n when we assume the underlying Gaussian processes are autoregressive.

AR models are useful to us in three ways. First, the process of generating data from the χ^2 distributions for simulations becomes significantly more efficient. Second, the structure of the inverse covariance matrix $\mathbf{Q}$ is greatly simplified, resulting in a k-step memory algorithm for computing the multivariate pdfs. It is this property that reduces the computational complexity from being exponential in n to being linear in n. Finally, the assumption of an AR model leads to a multivariate dependency structure that is defined by just a handful of parameters, a situation that is more favorable for parameter estimation.

The χ^2 random variables $\{Z_t\}$ are defined in terms of an underlying sequence $\{\mathbf{U}_t\}$ of Gaussian random vectors via (3.2). Generating data for simulations amounts to generating the Gaussian vectors with the appropriate covariance structure. The generation of the Gaussian variables may be done most efficiently if they are generated by a linear recursion,

$$\mathbf{U}_t + a_1\mathbf{U}_{t-1} + \cdots + a_k\mathbf{U}_{t-k} = b\xi_t \tag{3.17}$$

where $\xi_t = (\xi_{t1}, \ldots, \xi_{td})'$ is a vector of zero-mean, unit-variance Gaussian random variables. We assume that ξ_{si} is independent of ξ_{tj} for $s \neq t$ or $i \neq j$. The model defined by (3.17) is called an *autoregressive model* of order k. It is helpful to think of the vector process $\{\mathbf{U}_t\}$ as a collection of d identical scalar processes $\{U_{ti}\}$, $1 \leq i \leq d$. For simplicity in the following discussion, we consider a scalar AR process

$$U_t + a_1U_{t-1} + \cdots + a_kU_{t-k} = b\xi_t \tag{3.18}$$

and the results apply to each of the scalar processes that constitute $\{\mathbf{U}_t\}$.

The AR model defined by (3.18) involves the $k+1$ parameters $a_1, \ldots, a_k$, and b. Thus, in general, we may match at most $k+1$ given values of the autocorrelation function $R_U(\tau)$, defined by

$$R_U(\tau) = \mathrm{E}\, U_tU_{t+\tau}. \tag{3.19}$$

It is well known, in fact, that the kth-order AR model can match exactly the first $k+1$ values $R_U(0), R_U(1), \ldots, R_U(k)$. To do this, the model parameters are chosen to solve a system of linear equations known as the Yule-Walker equations [14],

$$\begin{bmatrix} R_U(0) & R_U(1) & R_U(2) & \cdots & R_U(k) \\ R_U(1) & R_U(0) & R_U(1) & \cdots & R_U(k-1) \\ R_U(2) & R_U(1) & R_U(0) & \cdots & R_U(k-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_U(k) & R_U(k-1) & R_U(k-2) & \cdots & R_U(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} b^2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3.20}$$

The Yule-Walker equations may be solved extremely efficiently using the well-known Levinson-Durbin algorithm. This algorithm uses a recursion to produce the model parameters for the kth-order AR model from the model parameters for the $(k-1)$th-order AR model and an additional value $R_U(k)$ of the autocorrelation function. Let $a_i(j), 1 \le i \le j$ and $b(j)$ denote the model parameters for the jth-order AR model. With $a_0(j) = 0$ for all j and $b(0) = R_U(0)$, the recursive formulas for the Levinson-Durbin algorithm at the jth step are

$$\begin{aligned} a_j(j) &= -b^{-2}(j-1)\Big[R_U(j) + \sum_{i=1}^{j-1} a_i(j-1) R_U(j-i)\Big] \\ a_i(j) &= a_i(j-1) + a_j(j) a_{j-i}(j-1) \quad i = 1, \ldots, j-1 \\ b(j) &= b(j-1)\sqrt{1 - a_j^2(j)}. \end{aligned} \tag{3.21}$$

There is also a backward recursion [36], whereby we may obtain the coefficients for the model orders $k-1, k-2, \ldots, 1$ from the coefficients for order k:

$$\begin{aligned} a_i(j-1) &= \frac{a_i(j) - a_j(j) a_{j-i}(j)}{1 - a_j^2(j)} \quad i = i, \ldots, j-1 \\ b(j-1) &= \frac{b(j)}{\sqrt{1 - a_j^2(j)}}. \end{aligned} \tag{3.22}$$

For our purposes, the most significant aspect of the Levinson-Durbin algorithm is that it provides a convenient factorization of the inverse covariance matrix $\mathbf{Q}$. Suppose $\mathbf{Q}^{-1}$ is the covariance matrix for the AR process defined by (3.18), and let $a_i(j)$, $1 \le i \le j$, and $b(j)$ be the parameters of the jth-order AR model, as given by the Levinson-Durbin algorithm (3.21) or (3.22). Then $\mathbf{Q}$ has the factorization

$$\mathbf{Q} = \mathbf{L}\mathbf{L}' \tag{3.23}$$

where $\mathbf{L}$ is the lower triangular matrix

$$\mathbf{L} = \begin{bmatrix} b^{-1} & & & & & & \\ a_1 b^{-1} & \ddots & & & & & \\ \vdots & \ddots & b^{-1} & & & & \\ a_k b^{-1} & & a_1 b^{-1} & b(k-1)^{-1} & & & \\ & \ddots & \vdots & \vdots & & \ddots & \\ & & a_k b^{-1} & a_{k-1}(k-1)b(k-1)^{-1} & \cdots & b(0)^{-1} \end{bmatrix}. \tag{3.24}$$

The elements below the kth subdiagonal in (3.24) are all zero. Thus $L_{ij} = 0$ if $j > i$ or $j < i - k$. This factorization is quite useful in reducing the amount of computation needed to compute the multivariate χ^2 pdf.

3.3 Computation of the Probability Density Functions

The complex integration required to evaluate the n-dimensional pdf for the central or non-central χ^2 distribution is prohibitively expensive for even moderately large n. In the general case, the computational complexity of a numerical integration method increases exponentially in n. Furthermore, there may be doubts concerning the inaccuracy of such a numerical method. We now consider assumptions that permit us to simplify the computation. Our main assumption, which is required for all the methods considered here, is that the underlying Gaussian processes are kth-order autoregressive. Thus we assume that the underlying Gaussian processes are of the form (3.17). This assumption permits us to evaluate the integral numerically with a complexity that increases only linearly in n.

3.3.1 First-Order Markov PDF

If the underlying Gaussian sequences are first-order autoregressive, then the matrix $\mathbf{Q}$ is three-banded, that is, $q_{ij} = 0$ for $|i - j| > 1$. This leads to significant reductions in the computation required to evaluate the multivariate χ^2 pdfs. In particular, Miller shows in [41, pp. 37–39] that the resulting $\chi^2(d)$ variables are Markovian. Thus, at most the bivariate

and univariate pdfs must be computed. For general d, we have for the n-dimensional pdf

$$\begin{aligned} f_z(\mathbf{z}) &= f_z^{(1)}(z_1) \prod_{t=2}^{n} \frac{f_z^{(2)}(z_{t-1}, z_t)}{f_z^{(1)}(z_{t-1})} \\ &= f_z^{(1)}(z_1) \prod_{t=2}^{n} f_{z_2|z_1}(z_t|z_{t-1}) \end{aligned} \tag{3.25}$$

where $f_z^{(1)}$ and $f_z^{(2)}$ are the univariate and bivariate pdfs, respectively, and $f_{z_2|z_1}$ is the conditional pdf of Z_2 given Z_1. Let the covariance matrix for adjacent Gaussian random variables be

$$\mathbf{Q}^{-1} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Then the general univariate and bivariate $\chi^2(d)$ pdfs are

$$\begin{aligned} f_z^{(1)}(x) &= \frac{x^{(d/2)-1} e^{-x/2}}{2^{d/2}\Gamma(d/2)} \\ f_z^{(2)}(x,y) &= \frac{(xy)^{(d-2)/4}}{2^{(d/2)+1}|\rho|^{(d/2)-1}(1-\rho^2)\Gamma(d/2)} \\ &\quad \times \exp\left\{-\frac{x+y}{2(1-\rho^2)}\right\} I_{(d/2)-1}\left\{\frac{|\rho|\sqrt{xy}}{1-\rho^2}\right\}. \end{aligned} \tag{3.26}$$

where I_ν is the modified Bessel function of the first kind of order ν. In particular, for $d = 2$ we have

$$\begin{aligned} f_z^{(1)}(x) &= \tfrac{1}{2} e^{-x/2} \\ f_z^{(2)}(x) &= \frac{1}{4(1-\rho^2)} \exp\left\{-\frac{x+y}{2(1-\rho^2)}\right\} I_0\left\{\frac{\rho\sqrt{xy}}{1-\rho^2}\right\}. \end{aligned} \tag{3.27}$$

For $d = 3$ we have

$$\begin{aligned} f_z^{(1)}(x) &= \sqrt{\frac{x}{2\pi}} e^{-x/2} \\ f_z^{(2)}(x) &= \frac{1}{2\pi|\rho|\sqrt{1-\rho^2}} \exp\left\{-\frac{x+y}{2(1-\rho^2)}\right\} \sinh\left\{\frac{|\rho|\sqrt{xy}}{1-\rho^2}\right\}. \end{aligned} \tag{3.28}$$

And finally, for $d = 4$ we have

$$\begin{aligned} f_z^{(1)}(x) &= \frac{x}{4} e^{-x/2} \\ f_z^{(2)}(x) &= \frac{\sqrt{xy}}{8|\rho|(1-\rho^2)} \exp\left\{-\frac{x+y}{2(1-\rho^2)}\right\} I_1\left\{\frac{\rho\sqrt{xy}}{1-\rho^2}\right\}. \end{aligned} \tag{3.29}$$

The logarithm of the pdf can be computed as the sum of one-step memory nonlinearities. Thus if we assume a pdf of the form (3.25) under hypotheses H_0 and H_1, the log-likelihood ratio can be computed as a sum of one-step memory nonlinearities.

For autoregressive sequences of order greater than one, numerical integration is required. Two possible approaches to computing the integral in (3.6) numerically seem feasible. One method is to perform the integration in each of the variables using a standard numerical quadrature method, such as Gaussian quadrature. This approach is discussed later. An alternative approach is to evaluate the integral via a Monte Carlo method.

3.3.2 Monte Carlo Integration

Theorem 3.1 provides the justification for Monte Carlo integration as a means to compute the multivariate pdf. This method involves generating realizations of the random vectors $\{\mathbf{Y}_t, 1 \le t \le n\}$ and then approximating the expection in (3.14) as a sample average.

We begin with $d = 2$; later we extend the result to $d > 2$. Let $\Theta_1, \ldots, \Theta_n$ be independent random variables uniformly distributed over $[0, 2\pi)$. Then we may rewrite (3.6) as an expectation,

$$f_z(\mathbf{z}) = 2^{-n}|\mathbf{Q}|\, \mathrm{E}\left[\exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sqrt{z_i z_j}\cos(\Theta_i - \Theta_j)q_{ij}\right\}\right].$$

To evaluate the integrals numerically using a Monte Carlo method, we generate realizations of the random variables $\Theta_1, \ldots, \Theta_n$ and then take a sample average. This yields

$$f_z(\mathbf{z}) \approx 2^{-n}|\mathbf{Q}|\frac{1}{M}\sum_{\ell=1}^{M}\exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sqrt{z_i z_j}\cos\left[\theta_i(\ell) - \theta_j(\ell)\right]q_{ij}\right\} \qquad (3.30)$$

where $\theta_i(\ell)$ denotes the ℓth realization of the random variable Θ_i.

The form given by (3.30) is not the most efficient way to evaluate the expectation. To improve the efficiency, we take advantage of the assumed autoregressive structure of the underlying Gaussian sequences. We begin with the form of the pdf given by (3.7) and (3.8). Since the Gaussian processes are kth-order autoregressive, the inverse covariance matrix $\mathbf{Q}$ is $(2k+1)$-banded and has the factorization given by (3.23) and (3.24). To evaluate the ℓth term of the sum in (3.30), we generate a realization $[\theta_1(\ell), \ldots, \theta_n(\ell)]$ of the $\{\Theta_t\}$ process and then form the sequences

$$u_t(\ell) = \sqrt{z_t}\cos\theta_t(\ell)$$

$$v_t(\ell) = \sqrt{z_t}\sin\theta_t(\ell).$$

Using the factorization (3.23) and setting $\bar{\mathbf{u}} = L'\mathbf{u}$ and $\bar{\mathbf{v}} = L'\mathbf{v}$, we have

$$\bar{\mathbf{u}}'\bar{\mathbf{u}} = \mathbf{u}'\mathbf{Q}\mathbf{u}$$

$$\bar{\mathbf{v}}'\bar{\mathbf{v}} = \mathbf{v}'\mathbf{Q}\mathbf{v}.$$

Note, however, that since the elements L_{ij} of the matrix $\mathbf{L}$ are 0 for $i - j > k$, we may compute $\bar{u}_t(\ell)$ and $\bar{v}_t(\ell)$ from $u_t(\ell), \ldots, u_{t-k}(\ell)$ and $v_t(\ell), \ldots, v_{t-k}(\ell)$ when $t > k$. In fact, the computation is done using a recursion involving the autoregressive parameters,

$$\begin{aligned} \bar{u}_t &= b^{-1} \sum_{j=0}^{k} a_j u_{t-j} \\ \bar{v}_t &= b^{-1} \sum_{j=0}^{k} a_j v_{t-j}. \end{aligned} \tag{3.31}$$

For $0 < t \leq k$, we may use formulas similar to (3.31),

$$\begin{aligned} \bar{u}_t &= b^{-1}(m) \sum_{j=0}^{m} a_j(m) u_{t-j} \\ \bar{v}_t &= b^{-1}(m) \sum_{j=0}^{m} a_j(m) v_{t-j} \end{aligned} \tag{3.32}$$

where $m = t$ and $b(m)$ and $a_j(m)$, $1 \leq j \leq m$ are the autoregressive parameters for the mth-order autoregressive model and $a_0(m) = 1$. Thus, we may consider (3.32) to be the general formula, with $m = \min(t, k)$.

Note that the pdf (3.7) may be written

$$f_z(\mathbf{z}) = 2^{-n}|\mathbf{Q}|(2\pi)^{-n} \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\left\{-\tfrac{1}{2} \sum_{i=1}^{n} (\bar{u}_i^2 + \bar{v}_i^2)\right\} d\theta_1 \cdots d\theta_n. \tag{3.33}$$

In (3.33), u_i and v_i are functions of $\theta_i, \ldots, \theta_{i-m}$ and $z_i, \ldots, z_{i-m}$, where $m = \min(k, i-1)$. Thus the Monte Carlo approximation becomes

$$f_z(\mathbf{z}) \approx 2^{-n}|\mathbf{Q}| \frac{1}{M} \sum_{\ell=1}^{M} \exp\left\{-\tfrac{1}{2} \sum_{i=1}^{n} \left[\bar{u}_i^2(\ell) + \bar{v}_i^2(\ell)\right]\right\}.$$

For the case of $d > 2$, we must generate a realization $\{\mathbf{y}_t, 1 \leq t \leq n\}$ of the rectangular coordinate vectors $\{\mathbf{Y}_t, 1 \leq t \leq n\}$ as defined in Theorem 3.1. The simplest way to generate a vector $\mathbf{Y}$ that is uniformly distributed on the unit hypersphere is to set

$$Y_i = \frac{W_i}{\sum_{j=1}^{d} W_j^2} \tag{3.34}$$

where the W_j are i.i.d. Gaussian random variables. The method defined by (3.34) is not the most computationally efficient, however. More efficient methods are given in [24, Section 7.1]. We provide here the methods for $d = 3$ and $d = 4$ that were derived by Marsaglia. For $d = 3$, generate independent deviates W_1 and W_2 that are uniform on $(-1, 1)$, rejecting them until $S = W_1^2 + W_2^2 < 1$. When the inequality is satisfied, set

$$Y_1 = 2W_1(1 - S)^{1/2}$$

$$Y_2 = 2W_2(1 - S)^{1/2}$$

$$Y_3 = 1 - 2S.$$

For $d = 4$, generate W_1 and W_2 uniform on $(-1, 1)$. Reject them until $S_1 = W_1^2 + W_2^2 < 1$. Similarly, generate W_3 and W_4 until $S_2 = W_3^2 + W_4^2 < 1$. Then set

$$Y_1 = W_1$$

$$Y_2 = W_2$$

$$Y_3 = W_3 \left(\frac{1 - S_1}{S_2} \right)^{1/2}$$

$$Y_4 = W_4 \left(\frac{1 - S_1}{S_2} \right)^{1/2}.$$

The accuracy of the Monte Carlo integration is measured in terms of the variance of the sample average. Unfortunately, the variance tends to be unacceptibly large for moderate values of M and decreases as $O(\sqrt{M})$. There are ways to reduce the variance, however. For example, we could break up the region over which the integration is performed into subregions and add together the approximate integrals taken over each subregion. To define a subregion, we may restrict each $\mathbf{Y}_t$ to take its values in a subregion of the unit hypersphere. The sum taken over each of the subregions could be computed using the idea that is discussed in the next subsection on Gaussian quadrature. This modified Monte Carlo method is similar to Gaussian quadrature, the difference being that the nodes for the Monte Carlo integration are chosen randomly.

3.3.3 Gaussian Quadrature

An alternative to using the Monte Carlo integration method is to use a numerical quadrature method and take advantage of the form of the inverse covariance matrix given

by (3.23) and (3.24) to simplify the computation. Because of the $(2k+1)$-banded structure of the matrix, the complexity of the numerical integration method is linear in n.

First consider the straightforward numerical integration of the multivariate $\chi^2(2)$ pdf given by

$$f_z^{(n)}(\mathbf{z}) = (4\pi)^{-n}|\mathbf{Q}| \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\left\{-\tfrac{1}{2}\sum_{s=1}^{n}\sum_{t=1}^{n}\sqrt{z_s z_t}\cos(\theta_s - \theta_t)q_{st}\right\} d\theta_1 \cdots d\theta_n. \quad (3.35)$$

Let $\theta_1, \ldots, \theta_N$ and $w_1, \ldots, w_N$ be, respectively, the nodes and weights of a Gaussian quadrature method such that

$$\int_0^{2\pi} f(x)\,dx \approx \sum_{i=1}^{N} f(\theta_i)\, w_i.$$

Letting I denote the integral in (3.35), we have

$$I \approx \sum_{i_n=1}^{N} \cdots \sum_{i_1=1}^{N} \exp\left\{-\tfrac{1}{2}\sum_{s=1}^{n}\sum_{t=1}^{n}\sqrt{z_s z_t}\cos(\theta_{t,i_t} - \theta_{s,i_s})q_{st}\right\} \prod_{t=1}^{N} w_{t,i_t}. \quad (3.36)$$

It can be seen from (3.36) the complexity of the "brute force" method is on the order of N^n; that is, it increases exponentially with n.

When the underlying Gaussian processes are kth-order autoregressive, then the pdf may be written in the form

$$f_z(\mathbf{z}) = (4\pi)^{-n}|\mathbf{Q}| \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\left\{-\tfrac{1}{2}\sum_{t=1}^{n}(\bar{u}_t^2 + \bar{v}_t^2)\right\} d\theta_1 \cdots d\theta_n \quad (3.37)$$

which is the same as (3.33). Here, $\bar{u}_t$ and $\bar{v}_t$ are functions of $\theta_t, \ldots, \theta_{t-k}$. Therefore, if we write $G(\bar{u}_i, \bar{v}_i)$ for $\exp\{-\frac{1}{2}(\bar{u}_i^2 + \bar{v}_i^2)\}$, then the integral in (3.37) may be expressed as

$$\begin{aligned} I &= \int_0^{2\pi} d\theta_n \cdots \int_0^{2\pi} d\theta_1 \prod_{t=1}^{n} G(\bar{u}_t, \bar{v}_t) \\ &= \int_0^{2\pi} d\theta_n \cdots \int_0^{2\pi} d\theta_{n-k} G(\bar{u}_n, \bar{v}_n) \int_0^{2\pi} d\theta_{n-k-1} G(\bar{u}_{n-1}, \bar{v}_{n-1}) \cdots \int_0^{2\pi} d\theta_1 \prod_{t=1}^{k+1} G(\bar{u}_t, \bar{v}_t) \end{aligned}$$

and the approximate integral (3.36) may be expressed as

$$\begin{aligned} I &\approx \sum_{i_n=1}^{N} w_{i_n} \cdots \sum_{i_1=1}^{N} w_{i_1} \prod_{t=1}^{n} G(\bar{u}_t, \bar{v}_t) \\ &= \sum_{i_n=1}^{N} w_{i_n} \cdots \sum_{i_{n-k}=1}^{N} w_{i_{n-k}} G(\bar{u}_n, \bar{v}_n) \cdots \sum_{i_1=1}^{N} w_{i_1} \prod_{t=1}^{k+1} G(\bar{u}_t, \bar{v}_t). \end{aligned} \quad (3.38)$$

Notice that $G(\bar{u}_t, \bar{v}_t)$ is a function of $i_t, \ldots, i_{t-m}$ for $m = \min(k, t-1)$. To show this explicitly, define

$$B_t(i_t, \ldots, i_{t-m}) = w_{i_t} \exp\left\{-\tfrac{1}{2}(\bar{u}_t^2 + \bar{v}_t^2)\right\}, \qquad m = \min(k, t-1). \tag{3.39}$$

Now, define $S_1, S_2, \ldots, S_{n-k+1}$ recursively by

$$S_1(i_k, \ldots, i_1) = \prod_{j=1}^{k} B_j(i_j, \ldots, i_1)$$

$$S_t(i_{t+k-1}, \ldots, i_t) = \sum_{i_{t-1}=1}^{N} B_{t+k-1}(i_{t+k-1}, \ldots, i_{t-1}) S_{t-1}(i_{t+k-2}, \ldots, i_{t-1}),$$

$$2 \le t \le n-k+1.$$

Finally, we have

$$I \approx \sum_{i_n=1}^{N} \cdots \sum_{i_{n-k+1}=1}^{N} S_{n-k+1}(i_n, \ldots, i_{n-k+1}).$$

At most N^{k+1} storage locations are required for the computation; however, the sums may be performed in a way so that only $2N^k$ strorage locations are necessary. The complexity of the algorithm is on the order of n; that is, the computation required increases linearly with n. This provides a significant time savings over the "brute force" method (3.36). A disadvantage of (3.38) is the large amount of storage required compared to (3.36), which requires minimal storage.

We may generalize this method to the case where $d > 2$. We need, of course, a numerical quadrature formula for integrating on the surface of a hypersphere. One possibilty for a quadrature formula may be derived from the spherical coordinates. A better possibility, though, is to use a formula that has been optimized specifically for integrating on the surface of a hypersphere [28]. Some formulas for $d = 3$ are given in Appendix C. For the general case, (3.39) is replaced by

$$B_t(i_t, \ldots, i_{t-m}) = w_{i_t} \exp\left\{-\tfrac{1}{2}(\bar{u}_{t1}^2 + \cdots + \bar{u}_{td}^2)\right\}, \qquad m = \min(k, t-1) \tag{3.40}$$

where

$$\bar{u}_{ti} = b^{-1}(m) \sum_{j=0}^{m} a_j(m) u_{t-j,i}$$

and

$$u_{ti} = y_i \sqrt{z_t}.$$

The y_i are the rectangular coordinates of the nodes of the quadrature formula.

The limiting factor for Gaussian integration is the amount of storage necessary for the algorithm, which is on the order of N^k, where N is the number of nodes per dimension and may be on the order of 100. For large k, Gaussian integration may be too expensive. The amount of storage also depends on the number of degress of freedom of the χ^2 distribution.

3.3.4 Examples

We now look at a few examples where we compare the different methods for computing the multivariate pdfs. We want to determine, if possible, just what we can expect in the accuracy of the numerical methods.

First we consider examples with $k = 1$. As shown in Section 3.3.1, we may, in this case, compute the exact value of the pdf without evaluating any integrals. Thus, we need not use an approximate numerical method. However, since we can compute the pdfs by using both the exact formula as well as one of several numerical methods, we may obtain a precise picture of how accurate the numerical methods are. We include this case as a first step in verifying that our approximate methods are valid. If comparison of the exact result and approximate results do not agree at this level, then we may have serious doubts concerning the value of the numerical approximations when $k > 1$. On the other hand, if the agreement between exact and approximate results is exceptional, then we may have much more confidence in using the approximate methods for $k > 1$.

The results for $d = 2$ and $d = 3$ are shown in Tables 3.1–3.6. The parameters shown in the tables are these: k is the order of the underlying AR Gaussian process, n is the dimensionality of the multivariate pdf computed, and N is the number of nodes in the numerical quadrature method. The value of ρ in (3.26) is taken to be $\rho = 0.9$. The pdfs were evaluated at a different set of values for each column; the values were generated by taking a random sample from the multivariate distribution. Tables 3.1 and 3.2 display the computed values of the pdf with $n = 10, 25, 50, 100$. The quadrature formulas used for $d = 2$ are those with nodes distributed symmetrically around the unit circle and uniform weights. The quadrature formulas for $d = 3$ have nodes symmetrically distributed on the surface of the unit sphere and nonuniform weights, as given in Appendix C. These quadrature formulas are optimal in the sense that they have the highest degree for a fixed number of nodes [28] (alternatively, they have the fewest number of nodes for a given degree). By

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Exact	9.4838e−06	1.5208e−15	4.2129e−17	1.6098e−49
Quadrature $N = 10$	1.0034e−05	2.7534e−15	4.9809e−17	8.1714e−48
Quadrature $N = 25$	9.4838e−06	1.5209e−15	4.2129e−17	1.6113e−49
Quadrature $N = 50$	9.4838e−06	1.5208e−15	4.2129e−17	1.6098e−49
Quadrature $N = 100$	9.4838e−06	1.5208e−15	4.2129e−17	1.6098e−49

Table 3.1: Values of the computed pdf for $k = 1$ and $d = 2$.

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Exact	1.1451e−05	8.7318e−16	7.9333e−20	1.0483e−51
Quadrature $N = 26$	1.6185e−05	9.0382e−15	2.1816e−19	2.4349e−45
Quadrature $N = 50$	1.1576e−05	1.2104e−15	8.2075e−20	1.4630e−50
Quadrature $N = 110$	1.1451e−05	8.7987e−16	7.9336e−20	1.1617e−51
Quadrature $N = 146$	1.1451e−05	8.7435e−16	7.9333e−20	1.0729e−51

Table 3.2: Values of the computed pdf for $k = 1$ and $d = 3$.

definition, a quadrature formula is said to be of degree m if it will yield an exact result for any polynomial of degree less than or equal to m.

Because of the extremely small magnitudes of the computed values, we also compare the logarithms of the computed values, as shown in Tables 3.3 and 3.4. Finally, we have the relative error of the logarithms of the computed values shown in Tables 3.5 and 3.6. One obvious conclusion that we may draw from these results is that there is no improvement for $d = 2$ when we go from 50 nodes to 100. Also, if we compare the columns in Tables 3.5 and 3.6, we see that sometimes the accuracy improves with increasing n. This result is explained by the fact that the data at which the pdf was evaluated is random and changes from column to column. Evidently, the accuracy obtained does depend on the particular point at which the pdf is evaluated.

Next, we look at the case of $k = 2$. Results corresponding to those in Tables 3.1–3.6 are shown in Tables 3.7–3.12. In this case, we are unable to obtain the exact result, and thus we lack the benchmark that is available when $k = 1$. The relative error shown in Tables 3.11 and 3.12 is taken with respect to the numerical quadrature method with $N = 100$ and $N = 146$ nodes, respectively. This is shown explicitly, since the last row contains only zeros. The first two correlation coefficients for the underlying Gaussian distribution are 0.9 and

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Exact	−1.1566e+01	−3.4120e+01	−3.7706e+01	−1.1235e+02
Quadrature $N = 10$	−1.1510e+01	−3.3526e+01	−3.7538e+01	−1.0842e+02
Quadrature $N = 25$	−1.1566e+01	−3.4120e+01	−3.7706e+01	−1.1235e+02
Quadrature $N = 50$	−1.1566e+01	−3.4120e+01	−3.7706e+01	−1.1235e+02
Quadrature $N = 100$	−1.1566e+01	−3.4120e+01	−3.7706e+01	−1.1235e+02

Table 3.3: Logarithm of the computed pdf for $k = 1$ and $d = 2$.

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Exact	−1.1378e+01	−3.4674e+01	−4.3981e+01	−1.1739e+02
Quadrature $N = 26$	−1.1031e+01	−3.2337e+01	−4.2969e+01	−1.0273e+02
Quadrature $N = 50$	−1.1367e+01	−3.4348e+01	−4.3947e+01	−1.1475e+02
Quadrature $N = 110$	−1.1378e+01	−3.4667e+01	−4.3981e+01	−1.1728e+02
Quadrature $N = 146$	−1.1378e+01	−3.4673e+01	−4.3981e+01	−1.1736e+02

Table 3.4: Logarithm of the computed pdf for $k = 1$ and $d = 3$.

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Quadrature $N = 10$	4.8755e−03	1.7397e−02	4.4414e−03	3.4954e−02
Quadrature $N = 25$	−5.9198e−09	9.0014e−07	−7.3134e−10	8.6190e−06
Quadrature $N = 50$	−5.9674e−09	1.2067e−09	−7.9645e−10	1.6380e−10
Quadrature $N = 100$	−5.9674e−09	1.2067e−09	−7.9645e−10	1.6380e−10

Table 3.5: Relative error of the log pdf for $k = 1$ and $d = 2$.

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Quadrature $N = 26$	3.0417e−02	6.7401e−02	2.3001e−02	1.2487e−01
Quadrature $N = 50$	9.6072e−04	9.4190e−03	7.7249e−04	2.2456e−02
Quadrature $N = 110$	7.8001e−07	2.2029e−04	7.2908e−07	8.7568e−04
Quadrature $N = 146$	3.8389e−08	3.8893e−05	3.8098e−08	1.9809e−04

Table 3.6: Relative error of the log pdf for $k = 1$ and $d = 3$.

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Quadrature $N = 10$	9.7810e−06	2.6446e−15	3.2299e−17	8.8031e−48
Quadrature $N = 25$	9.2228e−06	1.4358e−15	2.7225e−17	1.7379e−49
Quadrature $N = 50$	9.2228e−06	1.4357e−15	2.7225e−17	1.7358e−49
Quadrature $N = 100$	9.2228e−06	1.4357e−15	2.7225e−17	1.7358e−49

Table 3.7: Values of the computed pdf for $k = 2$ and $d = 2$.

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Quadrature $N = 26$	1.6217e−05	8.6876e−15	1.1772e−19	2.0769e−45
Quadrature $N = 50$	1.1443e−05	1.1482e−15	4.2788e−20	1.2832e−50
Quadrature $N = 110$	1.1305e−05	8.1147e−16	4.1273e−20	9.7910e−52
Quadrature $N = 146$	1.1305e−05	8.0496e−16	4.1272e−20	8.9658e−52

Table 3.8: Values of the computed pdf for $k = 2$ and $d = 3$.

0.8, so that $\mathrm{E}\, U_1 U_2 = 0.9$ and $\mathrm{E}\, U_1 U_3 = 0.8$.

Looking at the bottom two rows in Table 3.11, we see that the relative error between the methods with $N = 50$ and $N = 100$ is on the order of machine epsilon for double precision arithmetic. Thus, the results are the same to within round-off error. Considering the results obtained for $d = 2$ with both $k = 1$ and $k = 2$, it appears that $N = 50$ will yield sufficient accuracy.

For $d = 3$, however, exceptional accuracy is not so easy to obtain. With $N = 146$, the accuracy for computing the logarithm of the pdf is probably on the order of 10^{-3} or 10^{-4} for most data points when the dimensionality is $n = 100$. Just how much accuracy is needed for successful discrimination is not obvious. Given the possibility of statistical fluctuations, where the magnitude of the pdf may fluctuate by several orders of magnitude, the accuracy of 10^{-4} for the log-pdf, as observed in these results, may very well be more than sufficient. These answers are best obtained, however, by actually running simulations to test the discriminators. We take up this task in Section 3.6 of this chapter.

3.4 Parameter Estimation

Given our capability for computing the values of the multvariate pdfs, the method of maximum likelihood (ML) is certainly one option that is available to us for estimating the

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Quadrature $N = 10$	−1.1535e+01	−3.3566e+01	−3.7972e+01	−1.0835e+02
Quadrature $N = 25$	−1.1594e+01	−3.4177e+01	−3.8142e+01	−1.1227e+02
Quadrature $N = 50$	−1.1594e+01	−3.4177e+01	−3.8142e+01	−1.1228e+02
Quadrature $N = 100$	−1.1594e+01	−3.4177e+01	−3.8142e+01	−1.1228e+02

Table 3.9: Logarithm of the computed pdf for $k = 2$ and $d = 2$.

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Quadrature $N = 26$	−1.1030e+01	−3.2377e+01	−4.3586e+01	−1.0289e+02
Quadrature $N = 50$	−1.1378e+01	−3.4401e+01	−4.4598e+01	−1.1488e+02
Quadrature $N = 110$	−1.1390e+01	−3.4748e+01	−4.4634e+01	−1.1745e+02
Quadrature $N = 146$	−1.1390e+01	−3.4756e+01	−4.4634e+01	−1.1754e+02

Table 3.10: Logarithm of the computed pdf for $k = 2$ and $d = 3$.

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Quadrature $N = 10$	5.0685e−03	1.7874e−02	4.4802e−03	3.4969e−02
Quadrature $N = 25$	5.7453e−11	1.3627e−06	6.5001e−11	1.0804e−05
Quadrature $N = 50$	7.0144e−16	−3.0576e−16	−4.2369e−16	4.7924e−15
Quadrature $N = 100$	0e+00	0e+00	0e+00	0e+00

Table 3.11: Relative error of the log pdf for $k = 2$ and $d = 2$.

	$n = 10$	$n = 25$	$n = 50$	$n = 100$
Quadrature $N = 26$	3.1680e−02	6.8445e−02	2.3484e−02	1.2469e−01
Quadrature $N = 50$	1.0675e−03	1.0218e−02	8.0816e−04	2.2640e−02
Quadrature $N = 110$	8.9927e−07	2.3162e−04	7.4327e−07	7.4903e−04
Quadrature $N = 146$	0e+00	0e+00	0e+00	0e+00

Table 3.12: Relative error of the log pdf for $k = 2$ and $d = 3$.

parameters of the pdf. In addition to the fact that ML estimation has excellent asymptotic properties, it seems especially appropriate in our case since the pdfs will be used for LRT discriminators. We may maximize the multivariate pdf (or the log-pdf) by using a numerical optimization procedure for unconstrained multivariate optimization. Such methods, though, are outside of the scope of this dissertation.

On the other hand, we may go the same route that many other researchers have taken in the estimation of ARMA parameters: consider the possibility of estimating the parameters of the multivariate pdfs from sample moments. The rationale for this approach is the same as the rationale assumed in the estimation of ARMA parameters, namely, a lesser computational burden. In fact, since we assume that the underlying Gaussian processes in our model are autoregressive, the methods used in the estimation of AR parameters are especially relevant. Because we are able to write an explicit relationship between the moments of the χ^2 and the moments of the underlying Gaussian, we are able to adapt many of the methods for AR estimation to estimate the parameters for our models. The adaptation of these methods is the primary topic of this chapter.

Before we delve into the main topic, however, there are a few technical details that deserve our attention. First, we assume that the univariate pdf has already been estimated. Numerous methods are available for this, including both parametric and non-parametric methods. Whenever the observations are asymptotically independent, many such techniques for estimating the univariate distributions are essentially the same as the techniques used for independent observations.

Second, since we are proposing the use of transformation models, we have the option of using either the raw data $\{x_t, 1 \leq t \leq n\}$ or the transformed data $\{z_t, 1 \leq t \leq n\}$, where $z_t = \psi(x_t)$, in the estimation procedure. This issue is certainly relevant to the performance of the estimator. For our method, we assume that the transformed data z_t is used. The crucial issue is whether there exists an explicit functional relationship expressible in closed form between the estimated moments and the moments of the underlying Gaussian AR processes. If such a relationship exists between the moments of x_t and the moments of the Gaussian, then the method presented below may be adapted to permit use of the sample moments of x_t.

One final issue that we must deal with involves the determination of a model. This decision involves a choice between a transformed Gaussian model and a transformed $\chi^2(d)$ model. In the case that we choose a $\chi^2(d)$ model, we must also decide on a value for d. In

addition, we must determine the order k of the AR model. In deciding these issues, the fit of the model to the observations may not be the only criterion considered. Computational complexity of the multivariate pdf, for example, is also an important criterion. Considering that the pdfs are to be used in an LRT discriminator, the best method for choosing a model is perhaps maximum likelihood. All other things being equal, one should evaluate the multivariate pdf at the sample data for each model being considered, and then choose the model that yields the greatest value. Note that the number of models being considered should be fairly small, considering the limitations. For one thing, k cannot be too large or the computation required to compute the multivariate pdf may get out of hand. Also, d should not be too large, since the $\chi^2(d)$ distribution for large d tends to be more like the Gaussian. In the next section, we show that d should perhaps not be larger than 3. This latter fact is convenient in some ways, in that the computation required to compute the multivariate $\chi^2(d)$ requires numerical integration on the surface of a d-dimensional hypersphere, which is not a pleasant task for large d.

We now review some methods that have proven successful in the estimation of AR parameters. The Yule-Walker method, a straightforward, no-frills method, involves computing estimates $\hat{R}_U(0), \ldots, \hat{R}_U(k)$ of the autocorrelation function and then solving the Yule-Walker equations (3.20). It is well known that this method is asymptotically equivalent to the ML method for Gaussian AR processes [14]. Two methods are commonly used to estimate values of the autocorrelation function (3.19). The estimates obtained by

$$\hat{R}_U(\tau) = \frac{1}{N} \sum_{t=1}^{N-\tau} U_t U_{t+\tau} \tag{3.41}$$

are biased but are sometimes used because they guarantee that the resulting AR model will be stable. The alternative is

$$\hat{R}_U(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} U_t U_{t+\tau} \tag{3.42}$$

which is unbiased. Since the two methods are asymptotically equivalent, either method will suit our purpose if N is large.

All of the equations that constitute the Yule-Walker equations (3.20), with the exception of the first one, are of the form

$$a_1 R_U(t-1) + a_2 R_U(t-2) + \cdots + a_k R_U(t-k) = -R_U(t). \tag{3.43}$$

It is well known that equation (3.43) holds for kth order AR models for any $t > 1$. This fact provides the basis for the over-determined Yule-Walker method, where additional equations of the form (3.43) are added to the basic set (3.20) of Yule-Walker equations, and the resulting over-determined system is solved in the least squares sense. As Bruzzone and Kaveh show in [4], not all of the equations of the form (3.43) are equally important for all t; the equations with smaller values of t should be given more weight. Therefore, the over-determined Yule-Walker equations are best solved as a problem of weighted linear least squares. Specifically, we have the minimization problem

$$\text{minimize} \qquad (\mathbf{A}\theta - \mathbf{r})'\mathbf{W}(\mathbf{A}\theta - \mathbf{r}) \tag{3.44}$$

where $\theta = (a_1, \ldots, a_k)'$ is the vector of AR parameters, $\mathbf{r} = [-R_U(1), \ldots, -R_U(L)]'$,

$$\mathbf{A} = \begin{bmatrix} R_U(0) & R_U(1) & \cdots & R_U(k-1) \\ R_U(1) & R_U(0) & \cdots & R_U(k-2) \\ \vdots & \vdots & & \vdots \\ R_U(L-1) & R_U(L-2) & \cdots & R_U(L-k) \end{bmatrix} \tag{3.45}$$

and $\mathbf{W}$ is a symmetric positive definite weight matrix. The solution of the weighted least squares problem becomes the desired estimate $\hat{\theta} = (\hat{a}_1, \ldots, \hat{a}_k)'$. After the estimates of the parameters $(a_1, \ldots, a_k)$ are obtained, an estimate the gain b may be obtained by using the relation

$$b^2 = R_U(0) + a_1 R_U(1) + \cdots + a_k R_U(k) \tag{3.46}$$

where we substitute the estimates $\hat{R}_U(i)$ for the $R_U(i)$ and the components of $\hat{\theta}$ for the a_i.

The weight matrix $\mathbf{W}$ provides us with a mechanism for attaching more importance to some autocorrelation estimates than to others. Perhaps the best criterion for chosing $\mathbf{W}$ is the asymptotic covariance of $\hat{\theta}$, as suggested by Friedlander and Porat. In [12] and [13], they show that the asymptotic variance of the estimate $\hat{\theta}$, as a function of θ, is

$$\mathbf{V}(\theta) \stackrel{\text{def}}{=} \lim_{N\to\infty} \mathrm{E}(\hat{\theta} - \mathrm{E}\hat{\theta})(\hat{\theta} - \mathrm{E}\hat{\theta})' = (\mathbf{A}'\mathbf{W}\mathbf{A})^{-1}\mathbf{A}'\mathbf{W}\mathbf{D}\Sigma\mathbf{D}'\mathbf{W}\mathbf{A}(\mathbf{A}'\mathbf{W}\mathbf{A})^{-1} \tag{3.47}$$

where

$$\Sigma = \lim_{N\to\infty} \mathrm{E}(\hat{\mathbf{s}} - \mathbf{s})(\hat{\mathbf{s}} - \mathbf{s})'$$

is the asymptotic covariance matrix of the vector $\hat{\mathbf{s}} = [\hat{R}_Z(0), \ldots, \hat{R}_Z(L)]$ and $\mathbf{D}$ is a matrix whose ith column is

$$\mathbf{D}_i = \frac{\partial \mathbf{r}(\mathbf{s})}{\partial s_i} - \frac{\partial \mathbf{A}(\mathbf{s})}{\partial s_i}\theta. \tag{3.48}$$

A bound on $\mathbf{V}(\theta)$ is

$$\mathbf{V}(\theta) \geq \{\mathbf{A}'[\mathbf{D}\Sigma\mathbf{D}']^{-1}\mathbf{A}\}^{-1}, \tag{3.49}$$

and, as Friedlander and Porat show in [12], equality is achieved when the weight matrix $\mathbf{W}$ is taken as

$$\mathbf{W} = [\mathbf{D}\Sigma\mathbf{D}']^{-1}. \tag{3.50}$$

We interpret inequality in (3.49) to mean that the difference in the left-hand side and the right-hand side is positive semidefinite, which implies, in particular, that the diagonal elements of the matrix on the left are greater than or equal to the corresponding diagonal elements of the matrix on the right. Equation (3.50), therefore, provides the matrix $\mathbf{W}$ that yields the smallest asymptotic variance.

The problem with using (3.50) in practice is that it depends on the unknown parameter θ. Friedlander and Porat suggest in [13] that the following three step method be used.

1. Substitute the identity matrix I for the weight matrix W and solve the least squares problem for $\hat{\theta}^{(I)}$.
2. Compute the matrix $\hat{D}$ by using $\theta^{(I)}$.
3. Compute $\hat{W} = [\hat{D}\hat{\Sigma}\hat{D}]^{-1}$, where $\hat{\Sigma}$ is an estimate of Σ obtained from the observations.
4. Use $\hat{W}$ as the weight matrix to solve the least squares problem for $\hat{\theta}$.

To adapt this method to our problem of estimating the parameters of the underlying Gaussian processes, we need an explicit closed form functional relationship between the moments of the $\chi^2(d)$ process Z_t and the moments of the underlying Gaussian process U_t. Let R_Z denote the autocorrelation function for the $\chi^2(d)$ process and R_U the autocorrelation function for the underlying Gaussian processes. These are defined by

$$\begin{aligned} R_Z(\tau) &= \mathrm{E}\, Z_t Z_{t+\tau} \\ R_U(\tau) &= \mathrm{E}\, U_t U_{t+\tau}. \end{aligned} \tag{3.51}$$

The relationship between the two autocorrelation functions may be shown to be

$$R_Z(\tau) = d^2 + 2dR_U^2(\tau). \tag{3.52}$$

From (3.52) we may also derive the relations

$$\mathrm{Cov}[Z_t, Z_{t+\tau}] = 2dR_U^2(\tau)$$

and

$$\gamma_Z(\tau) \stackrel{\text{def}}{=} \frac{\mathrm{Cov}[Z_t, Z_{t+\tau}]}{\mathrm{Var}\, Z_t} = R_U^2(\tau). \tag{3.53}$$

Now we have from (3.43) and (3.53) that

$$a_1\sqrt{\gamma_Z(t-1)} + a_2\sqrt{\gamma_Z(t-2)} + \cdots + a_k\sqrt{\gamma_Z(t-k)} = -\sqrt{\gamma_Z(t)}. \tag{3.54}$$

This result leads us consider the least squares problem (3.44) where $\mathbf{r} = [-\sqrt{\gamma_Z(1)}, \ldots, -\sqrt{\gamma_Z(L)}]'$ and

$$\mathbf{A} = \begin{bmatrix} \sqrt{\gamma_Z(0)} & \sqrt{\gamma_Z(1)} & \cdots & \sqrt{\gamma_Z(k-1)} \\ \sqrt{\gamma_Z(1)} & \sqrt{\gamma_Z(0)} & \cdots & \sqrt{\gamma_Z(k-2)} \\ \vdots & \vdots & & \vdots \\ \sqrt{\gamma_Z(L-1)} & \sqrt{\gamma_Z(L-2)} & \cdots & \sqrt{\gamma_Z(L-k)} \end{bmatrix} \tag{3.55}$$

The rest of the problem is as before. The optimal weight matrix $\mathbf{W}$ is again given by (3.50), where $\mathbf{D}$ is given by (3.48).

Estimation methods based on sample moments are valuable in that they permit us to estimate the parameters when k or d is too large to allow accurate computation of the multivariate pdf. Thus, we may use the method to estimate parameters for use in simulations, where we need only to generate data and not to compute the pdf.

3.5 A Comparison of the PDFs

When using a transformation model with either a Gaussian or a χ^2 distribution as the underlying distribution, we may always match the univariate marginal pdf and the correlation of the distribution being modeled. Thus, to notice any differences in the different transformation models, we must look beyond the univariate pdfs and the second-order statistics. Possibly, we could look at higher-order statistics. Instead, we look at plots of bivariate pdfs so that we may visually compare the models and determine their differences. To compare the bivariate pdfs $f_1^{(2)}$ and $f_2^{(2)}$, we apply a transformation to map them into bivariate pdfs $\hat{f}_1^{(2)}$ and $\hat{f}_2^{(2)}$ that have the same marginal pdfs. In addition, we choose the correlation parameters of $f_1^{(2)}$ and $f_2^{(2)}$ so that $\hat{f}_1^{(2)}$ and $\hat{f}_2^{(2)}$ have the same correlation coefficient.

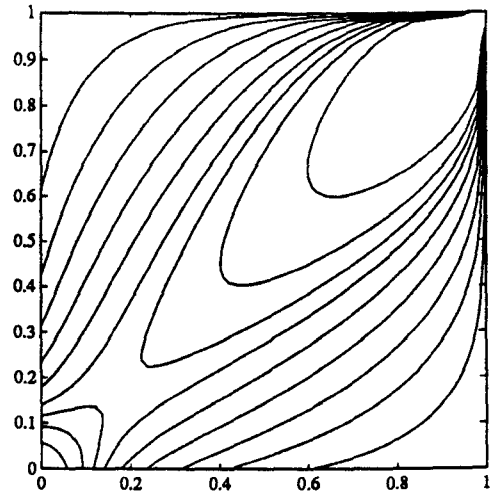


Figure 3.1: Contour plot of bivariate uniform pdf obtained by transformation of a $\chi^2(1)$ distribution.

First, we consider the case where $\hat{f}_1^{(2)}$ and $\hat{f}_2^{(2)}$ have uniform (0,1) marginals. The uniform marginals are convenient in that they permit us to view the whole of the pdf, they are symmetric, and they perhaps show most clearly the differences in the dependency structures. Contour plots of the bivariate uniform pdfs are shown in Figures 3.1–3.5. The contour plots are much better than surface plots for comparing the pdfs, and it is not hard to interpret them once we understand the basic shape of the pdfs. Since the random variables are positively correlated, the pdf takes its smallest values in the lower right-hand and upper left-hand corners. Likewise, it takes its largest values in the lower left-hand and upper right-hand corners. The underlying distribution for each of these plots are, respectively, the $\chi^2(1)$, $\chi^2(2)$, $\chi^2(3)$, $\chi^2(4)$, and Gaussian. In each case, the correlation coefficient of the uniform variables is 0.8.

The most radical differences are seen Figures 3.1 and 3.5, corresponding to the $\chi^2(1)$ and the Gaussian. The other plots seem to make a noticeable transition between the two extremes. The fact that this transition occurs stems from the fact that the $\chi^2(d)$ distribution, arising from a sum of i.i.d. random variables, tends to a Gaussian distribution under the central limit theorem. This tendency to a Gaussian holds for multivariate pdfs as well, according to the multivariate central limit theorem. There is enough similarity in the

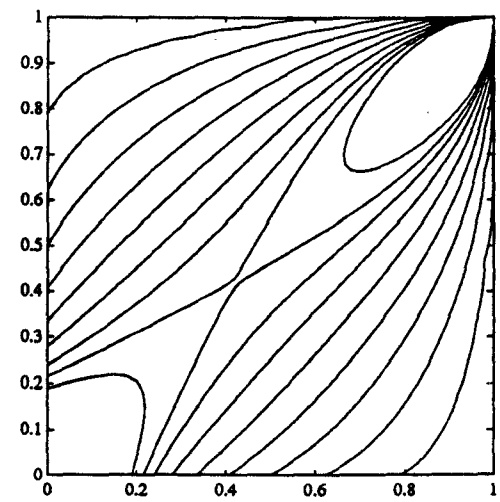


Figure 3.2: Contour plot of bivariate uniform pdf obtained by transformation of a $\chi^2(2)$ distribution.

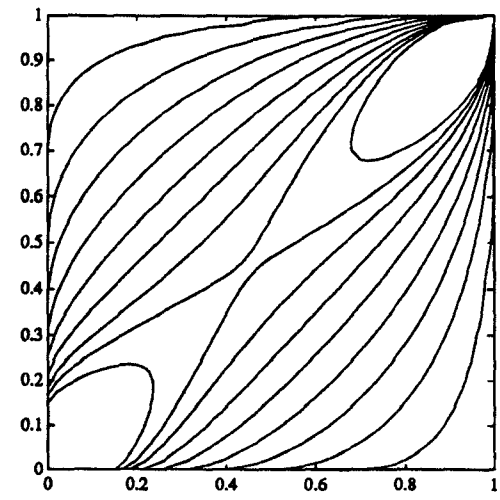


Figure 3.3: Contour plot of bivariate uniform pdf obtained by transformation of a $\chi^2(3)$ distribution.

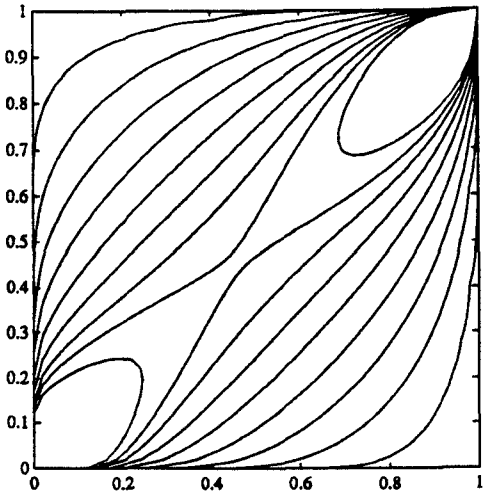


Figure 3.4: Contour plot of bivariate uniform pdf obtained by transformation of a $\chi^2(4)$ distribution.

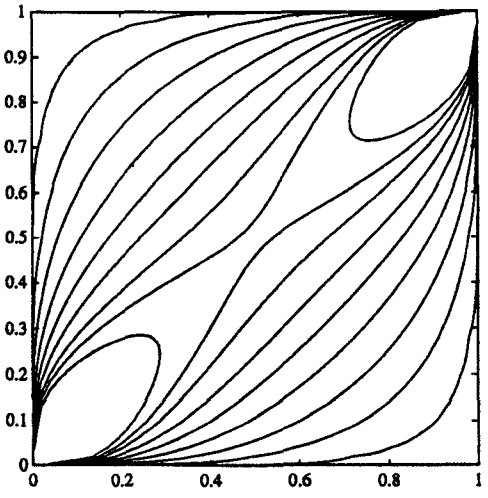


Figure 3.5: Contour plot of bivariate uniform pdf obtained by transformation of a Gaussian distribution.

plots corresponding to the $\chi^2(3)$ and $\chi^2(4)$ distributions to assume that they may be used interchangeably. One should perhaps choose between these two on the basis of convenience. Considering their similarity to the Gaussian, though, it is perhaps not wise to choose either of these in favor of the Gaussian, since they are both considerably more computationally expensive.

Note that any of these plots may be rotated by a multiple of 90 degrees to obtain other bivariate uniform pdfs. For example, if U_1 and U_2 are uniform random variables having a joint pdf shown in one of the figures, then the random variables $W_1 = 1 - U_1$ and $W_2 = 1 - U_2$ are also uniform random variables having a joint pdf that looks like the joint pdf of U_1 and U_2 rotated by 180 degrees. Thus, as there is a transition from the pdf in Figure 3.1, which is skewed to the upper right corner, to the symmetric pdf in Figure 3.5, we may continue this transition to a pdf that is skewed to the lower left corner by doing 180 degree rotations.

To see how the pdfs are apt to look for positive valued random variables, we also consider bivariate pdfs with exponentially distributed marginals. As before, the bivariate pdfs all have the same correlation coefficient, which is 0.8. Shown in Figures 3.6–3.8 are the bivariate exponential pdfs obtained by transformation of $\chi^2(1)$, $\chi^2(2)$, and Gaussian distributions. These plots are perhaps harder to interpret than the plots of the uniform pdfs. However, we may see some noticeable differences in the region near the origin. Since the pdfs all have the same correlation coefficient, we may find it useful in comparing their differences to think in terms of local correlation. In this context, the pdf in Figure 3.6, corresponding to the $\chi^2(1)$, seems to show the greatest local correlation near the origin, as compared to the other pdfs. We draw this conclusion from the fact that the contour lines appear the most sharply pointed at the origin in Figure 3.6. The pdf in Figure 3.7, which corresponds to the $\chi^2(2)$, seems to show the least local correlation near the origin.

3.6 Simulation Results

Finally, we evaluate the usefulness of the transformation models we have discussed for discrimination. The obvious discriminator to use is the likelihood ratio test (LRT). In this section we investigate the performance of such discriminators through simulations.

The first set of simulations are included to show the performance difference that one can expect when a $\chi^2(2)$ (exponential) distribution is chosen instead of a Gaussian

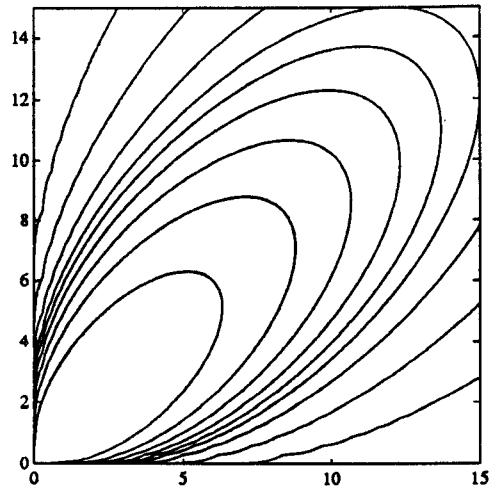


Figure 3.6: Contour plot of bivariate exponential pdf obtained by transformation of a $\chi^2(1)$ distribution.

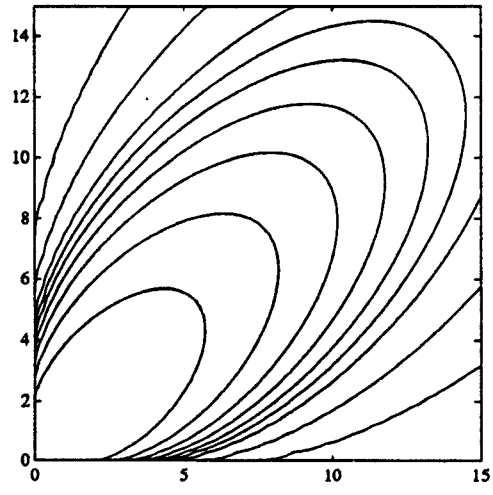


Figure 3.7: Contour plot of bivariate exponential pdf obtained by transformation of a $\chi^2(2)$ distribution.

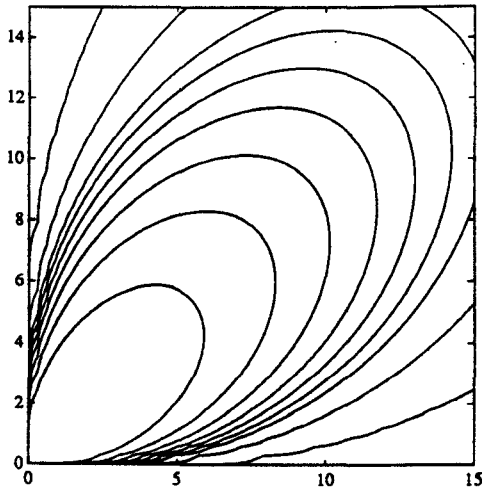


Figure 3.8: Contour plot of bivariate exponential pdf obtained by transformation of a Gaussian distribution.

for the underlying process of a transformation model. The distributions of the observed processes are Rayleigh and lognormal under the hypotheses H_0 and H_1, respectively. These distributions are used often in the modeling of radar signals, and therefore this set of simulations represents a typical real-world application. To obtain multivariate pdfs for the two signals, we may choose between any of the transformation models proposed in the earlier sections of this chapter. These include, namely, the Gaussian and the $\chi^2(d)$, where $d \leq 3$. In our simulations, the actual distributions of the signals that drive the simulations are obtained by a transformation of exponential variables. The exponential distribution is a natural choice for an underlying distribution in radar modeling, since the observed signal is often a function of the envelope of a complex signal. The popularity of the Weibull distribution in radar modeling attests to this fact.

Under both H_0 and H_1, we generated the observations for the simulation by first generating two first-order Gaussian AR processes U_{1t} and U_{2t}. We combined these to form an exponential process $Z_t = U_{1t}^2 + U_{2t}^2$. Finally, we obtained the observations by passing the exponential variables through a memoryless nonlinearity, $X_t = \psi^{-1}(Z_t)$. Under H_0, the observations have a Rayleigh distribution with relatively weak correlation. The adjacent correlation of the Gaussian variables—that is, the correlation between U_{1t} and U_{2t}—is

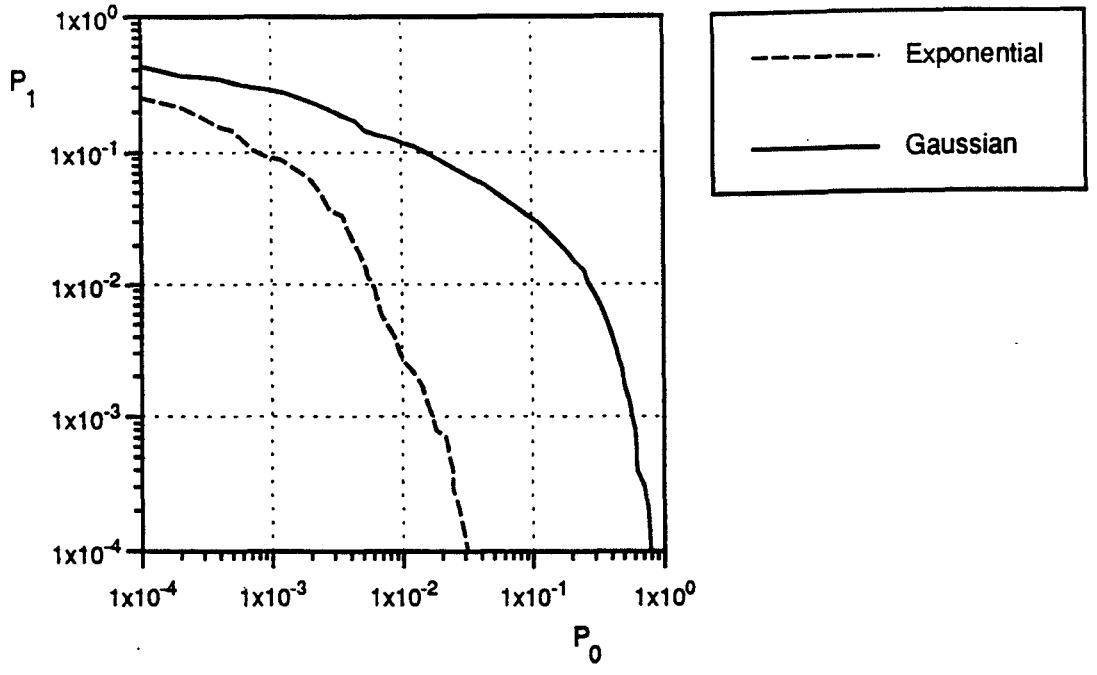


Figure 3.9: Plot of the error probabilities for lognormal vs. Rayleigh distributions, based on LRT for transformed Gaussian and transformed $\chi^2(2)$. Sample size is $n = 10$.

$\rho_0 = e^{-1/\tau_0}$ with $\tau_0 = 13$. Under H_1, the observations have a lognormal distribution, and the Gaussian variables have an adjacent correlation of $\rho_1 = e^{-1/\tau_1}$, with $\tau_1 = 130$.

The discriminator used is the LRT with $n = 10$ samples. After 1000 trials, we generated the plots shown in Figure 3.9. The curve labeled "Exponential" corresponds to the LRT that is derived under the assumption that the underlying processes are exponential. The correlation parameters for the pdfs in this case are exactly matched to those of the observations. The curve labeled Gaussian corresponds to the LRT derived under the assumption that the underlying processes are Gaussian. Thus, this LRT is the one that an analyst might use primarily for the sake of convenience, since the transformed Gaussian is a fairly simple distribution. The parameters of the Gaussian were chosen to match as closely as possible the true distribution of the observations. We note, however, that there is a mismatch in the autocorrelation values of the pdfs used in the LRT and the autocorrelation values of the observations. To match them exactly is an inconvenience that few analysts would endure. Therefore, we assume that the underlying Gaussian processes from

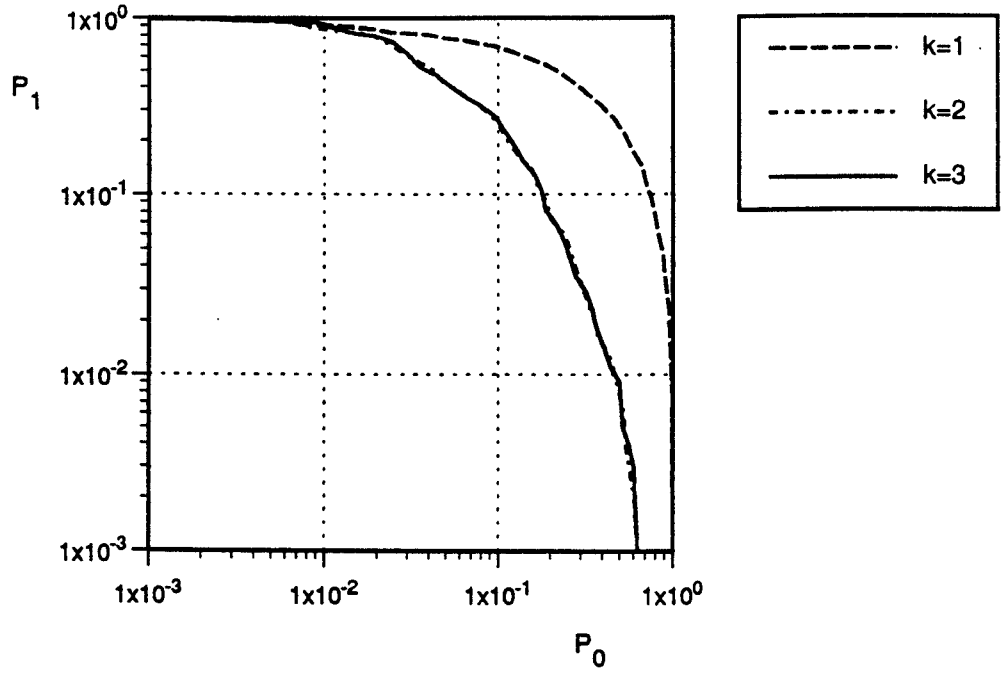


Figure 3.10: Plot of the error probabilities for LRT for exponential random variables with different AR order k. Sample size is $n = 20$. Number of nodes if $N = 20$.

which the multivariate pdfs are derived are again first-order autoregressive. We chose the adjacent correlation of the underlying Gaussian variables so that the adjacent correlation of the observed processes, both assumed and actual, were matched. This seems a reasonable thing to do given the small sample size of the test. The results of the simulations shown in Figure 3.9 indicate that there is clearly an advantage in implementing an LRT based on an underlying exponential distribution instead of an underlying Gaussian distribution. In the case of $k = 1$, the additional computational effort required is also minimal.

In the next set of simulations, we investigate the performance improvement that might be expected if we take a larger value for the AR order k or the number of nodes N. Concerning N, the question might be phrased differently, as we ask "How small of a value for N can we get by with?" This is important, since the computational expense increases with increasing N. While this is also true for k, there may be certain physical interpretations involving k that require that k not be too small. For example, if we are inclined to think that the AR processes have several distinct poles, then we would want to

keep k large enough to account for at least the most significant poles.

We assume now that the signals under H_0 and H_1 are exponential, with the same univariate distributions. This will permit us to perform controlled experiments, in which individual effects can be singled out. In this case, we are interested in the differences in correlation in the two signals. We assume that the underlying Gaussian variables U_{1t} and U_{2t}, from which the exponential variables are obtained, are fourth-order AR. We characterize the AR parameters in terms of their pole placements. Under H_0, we assume two complex pole pairs having magnitudes 0.9 and 0.7 and phases $\pi/8$ and $\pi/3$, respectively. Under H_1, we assume two complex pole pairs having magnitudes 0.9 and 0.2 and phases $\pi/8$ and $\pi/3$, respectively. Thus, both signals have their first pole pair in common, but the second pole pair has receded toward the origin under H_1. It is of interest to see, therefore, how well we can determine the prominence of the second pole, as under H_0, or its recession, as under H_1. In our simulations, we keep all other things equal, but change the value of k in the pdfs. The parameters other than k are matched to their true values in the simulations.

Results are shown for two different sample sizes in Figures 3.10 and 3.11. Figure 3.10 shows the results for $n = 20$ samples, with plots for $k = 1$, 2, and 3. In both cases, there is a significant improvement when we increase k from 1 to 2. There is virtually no improvement at all for a further increase in k from 2 to 3. These results, except for the ones with $k = 1$, were obtained by using Gaussian quadrature with $N = 20$ uniform nodes and weights to compute the multivariate pdfs. Figure 3.11 shows the results for $n = 40$ samples. Again, we reach the same conclusions.

Last of all, we consider the possibility of reducing the computational burden by reducing the number of nodes for the Gaussian integration. The results shown in Figure 3.12 as for the same situation as in Figure 3.11, except that the number of nodes is $N = 10$ instead of $N = 20$. For the case of $k = 3$, this is a reduction in computational expense of about 2^3, since the number of nodes is reduced by a factor of 2. Since there is no Gaussian integration when $k = 1$, the curve for $k = 1$ is identical in the two figures. Comparing the other curves, we see that there is not much difference in performance. The curve for $k = 2$ in Figure 3.12 is slightly better than the curve for for $k = 3$. Because of the tendency for error to be multiplied in iterated integrals, this may be evidence of a loss of accuracy. All things considered, one may decide to use the smaller number of nodes and sacrifice a little in performance for the sake of easier computation.

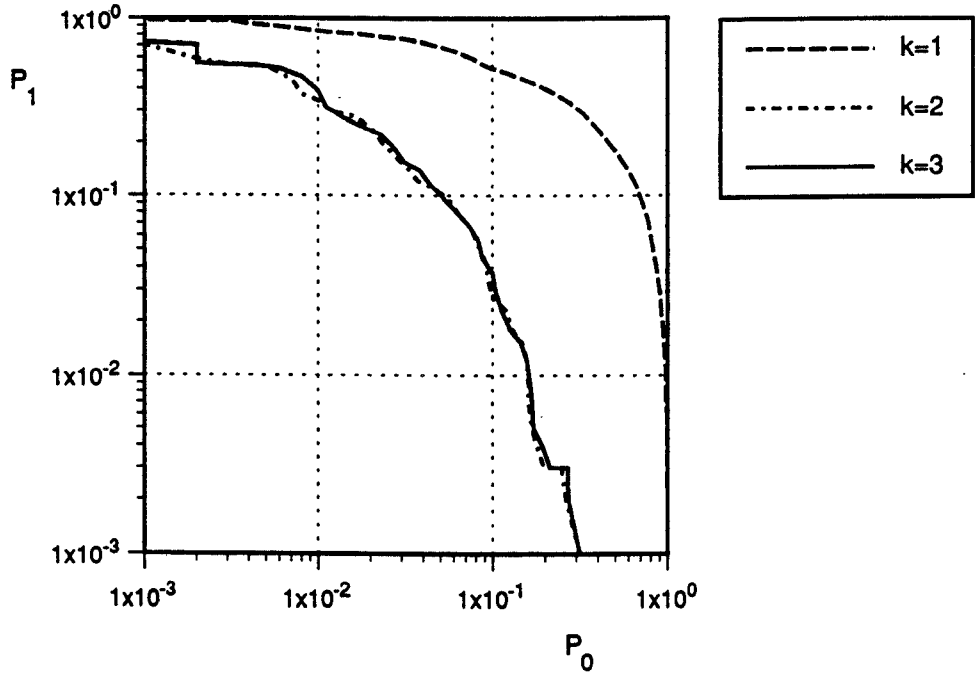


Figure 3.11: Plot of the error probabilities for LRT for exponential random variables with different AR order k. Sample size is $n = 40$. Number of nodes if $N = 20$.

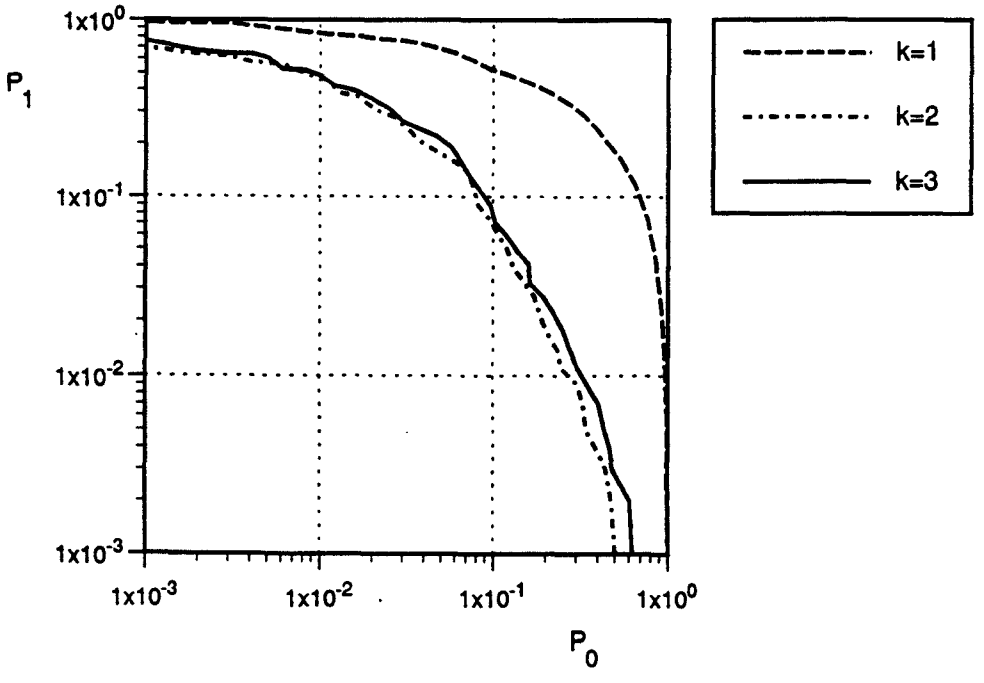


Figure 3.12: Plot of the error probabilities for LRT for exponential random variables with different AR order k. Sample size is $n = 40$. Number of nodes is $N = 10$.

3.7 Discussion

To summarize the main results of this chapter, we have seen how the multivariate χ^2 can be used as an alternative to the Gaussian in a transformation model. The use of the $\chi^2(2)$ seems especially appropriate for use in radar when the signal to be modeled is the envelope of a complex signal. The approximate methods for computing the multivariate χ^2 pdf can be quite accurate, although extreme accuracy is not required for LRT discriminators to be effective. A least squares method for estimating the parameters of ARMA models from sample moments can be adapted to estimate the parameters of the χ^2 models. Using this method, it is not necessary to compute the multivariate pdf; therefore, one may use this method to estimate the parameters for simulation studies where the computation of the multivariate pdf is not required. A comparison of contour plots of the different models reveals that there is a significant difference among them. This fact is corroborated by the simulation results, where we see a case where the χ^2 model yields much better results than the Gaussian model.

Although the emphasis in this chapter has been primarily discrimination using the LRT discriminator, there are other applications in which the theory presented here may be useful. One application is computer simulation. Transformed Gaussian processes are frequently used in simulations where correlated processes with specified univariate pdfs are needed. Certainly the χ^2 process should be considered as an alternative to the Gaussian in simulations, since the additional computation required is minimal. This is especially true when the processes generated for the simulations are positive valued. Another important application of these results is in the technique of importance sampling. This is a technique for reducing the number of simulations that must be performed in order to accurately estimate the probabilities of certain events. The technique requires that one be able to compute the multivariate pdf of the distribution from which the data is generated. As a specific example, we might consider the simulations for the one-step memory discriminators in Chapter 2. If the processes that drive the simulations are transformed χ^2 processes, then we may use the methods of this chapter to compute the multivariate pdfs, as required for importance sampling, and thereby significantly reduce the number of simulations needed to estimate the error probabilities.

Chapter 4

Hidden Markov Models

4.1 Introduction

We continue in this chapter on the theme of using an LRT discriminator. Therefore, our focus is again on obtaining flexible and computable multivariate probability density functions (pdfs). As in Chapter 3, we address the problem of obtaining multivariate pdfs from specified univariate pdfs. The solution to this problem that we develop here has both advantages and disadvantages compared to the solution considered in Chapter 3. On the advantage side, the hidden Markov models considered in this chapter are more flexible than the transformed χ^2 models. For instance, they are equally applicable to one-sided and two-sided processes. In addition, we may compute the exact value of the multivariate pdf with a minimal computational effort, unlike the χ^2 models where we compute only an approximation to the multivariate pdf. On the disadvantage side, if we choose to employ the most flexible model possible, then we have to deal with a large number of parameters, and therefore parameter estimation can become a problem. It is helpful to think of these hidden Markov models as partially parametric and partially nonparametric. The univariate pdfs are parametric, while the multivariate pdfs, constructed from the univariate pdfs in a structure involving a large number of parameters, are nonparametric in nature. We discuss methods to reduce the number of parameters involved, so that the resulting multivariate pdfs are more like parametric models, but in doing so, we reduce somewhat the flexibility of the models.

Hidden Markov models (HMMs) are especially useful for modeling certain nonstationary processes. Speech signals are a good example of such nonstationary processes, and

HMMs have been used successfully in automatic speech recognition [33], and other areas (see [9], [48], and the references therein). Much of the success of HMMs in speech recognition depends on the re-estimation method of Baum and others [2] for identifying the parameters of the models. There are, however, some problems with the method, as pointed out in [43] and [33]. There are ways around these problems, but they are a nuisance nonetheless. More recently, Ephraim, Dembo, and Rabiner [9] have proposed a minimum discrimination information method for identifying the parameters for a HMM.

The HMMs that we propose here are radically different from the "traditional" HMMs. The major difference concerns the output pdfs, or the conditional pdfs of the observation variables given the values of the underlying state variables. The most commonly used HMMs assume the that output distributions are either discrete probability mass functions or Gaussian pdfs. Some variants have been considered, however, that propose more general continuous output pdfs, including spherically symmetric distributions [35] and distributions belonging to the exponential family [39]. In the traditional HMMs, the emphasis appears to be on modeling nonstationary processes, with concern primarily for the conditional pdf of the observable variables. In other words, the underlying state sequence has some physical interpretation as a sequence that tracks the state of the process being modeled. In our models, however, the emphasis is on stationary processes with known unconditional output pdfs. The underlying state sequence has no physical interpretation other than that of providing a dependency structure in the multivariate pdf.

We begin this chapter by deriving the multivariate pdfs of our HMMs. This is followed be a discussion of the forward-backward algorithm for efficiently computing the multivariate and its gradient. Next, we look at some examples of "parsimonious" models, where a structure is imposed on the models so as to reduce the number of free parameters. Then parameter estimation is considered, and three different methods are presented. Finally, we present some simulation results, where we can see how well the methods work for discrimination.

4.2 Derivation of the Multivariate Distributions

Let $\{X_1, \ldots, X_N\}$ denote a collection of samples from the observation process. We make the assumption that the univariate marginal distribution for each sample is the same, and denote common cumulative distribution function (cdf) and probability density function

(pdf) by F_x and f_x, respectively. In this section, we derive a multivariate pdf for the samples that is compatible with the univariate pdf f_x. The basic idea for the construction depends on the following lemma.

Lemma 4.1 *Let $\{Y_1, \ldots, Y_M\}$ be independent random variables each having the cdf F, and let $\{Y^{(1)}, \ldots, Y^{(M)}\}$ denote their order statistics, that is, $Y^{(1)} \leq Y^{(2)} \leq \cdots \leq Y^{(M)}$ are same variables as $\{Y_1, \ldots, Y_M\}$ arranged in increasing order. Let Y be chosen randomly from among $\{Y^{(1)}, \ldots, Y^{(M)}\}$ with each of the M choices being equally likely. Then Y has the cdf F.*

Proof. Consider the event $\{Y^{(k)} \leq x\}$. This event occurs only if at least k of the variables $Y_1, Y_2, \ldots, Y_M$ take values less than x. This event can also be represented as the disjoint union of the events {exactly j of the Y_i's are $\leq x$} for $j = k, k+1, \ldots, M$. These latter events have a binomial probability distribution with probability of success being $F(x)$. Thus we have

$$P\{Y^{(k)} \leq x\} = \sum_{j=k}^{M} \binom{M}{j} F(x)^j \left[1 - F(x)\right]^{M-j}$$

Let S be the selection random variable, which is distributed uniformly on the set $\{1, 2, \ldots, M\}$. In accordance with the hypothesis of the theorem, we assume that S randomly selects one of the random variables $Y^{(1)}, Y^{(2)}, \ldots, Y^{(M)}$ to yield the value for Y. We now condition $P\{Y \leq x\}$ on the value of S and apply the law of total probability to

obtain

$$
\begin{aligned}
P\{Y \le x\} &= \sum_{k=1}^{M} P\{Y \le x | S = k\} P\{S = k\} \\
&= \frac{1}{M} \sum_{k=1}^{M} P\{Y^{(k)} \le x\} \\
&= \frac{1}{M} \sum_{k=1}^{M} \sum_{j=k}^{M} \binom{M}{j} F(x)^j \left[1 - F(x)\right]^{M-j} \\
&= \frac{1}{M} \sum_{k=1}^{M} k \binom{M}{k} F(x)^k \left[1 - F(x)\right]^{M-k} \\
&= \sum_{k=1}^{M-1} \frac{k}{M} \binom{M}{k} F(x)^k \left[1 - F(x)\right]^{M-k} + M \cdot \frac{1}{M} F(x)^M \\
&= F(x) \sum_{j=0}^{M-2} \binom{M-1}{j} F(x)^j \left[1 - F(x)\right]^{M-1-j} + F(x)^M \\
&= F(x)\left[1 - F(x)^{M-1}\right] + F(x)^M \\
&= F(x)
\end{aligned}
$$

Q.E.D.

To see how we use the lemma to construct the multivariate pdf, suppose we have a state process $\{S_1, S_2, \ldots, S_N\}$ where each random variable in the state process takes values in $\{1, 2, \ldots, M\}$ with equal probability; thus, S_t has a discrete uniform distribution on $\{1, 2, \ldots, M\}$ for each t. Now suppose we generate M independent values $Y_1, Y_2, \ldots, Y_M$ of a random variable Y that has the pdf f_x. We sort the values to obtain the order statistics $Y^{(1)} \le Y^{(2)} \le \ldots \le Y^{(M)}$ and use the value of S_1 to select one of these order statistics as the observation variable. Thus we set $X_1 = Y^{(S_1)}$. Since S_1 is uniformly distributed on the set $\{1, 2, \ldots, M\}$, by Lemma 4.1 X_1 has the pdf f_x. The remaining observation variables are generated by repeating the process. To generate X_t, we generate a new set of order statistics $Y^{(1)} \le Y^{(2)} \le \ldots \le Y^{(M)}$ and set X_t to the one selected by S_t. According to Lemma 4.1, each of the observation variables $X_1, X_2, \ldots, X_N$ generated in this way has the univariate pdf f_x.

If the set of Y-variables generated at time t_1 is independent of the set generated at time t_2, then the given the values of S_{t_1} and S_{t_2}, the random variables X_{t_1} and X_{t_2} are conditionally independent. However, the random variables $X_1, X_2, \ldots, X_N$ are not

unconditionally independent if the state variables are dependent. The dependency structure of the observation variables derives from the dependency structure of the state variables. This provides us with a flexible means to specify a dependency structure for the observation variables while maintaining their common univariate distribution. The process of moving from the state process to the observation process stochastically smooths and shapes the discrete distribution into the desired continuous distribution. However, the smoothing is not sufficient to destroy the dependency among the state variables as seen in the observation variables.

Let us now explicity construct the multivariate pdf of $\{X_1, \ldots, X_N\}$. Let $f_x^{[k]}$ denote the pdf of the k-th order statistic of a sequence of M independent and identically distributed (i.i.d.) random variables drawn from the pdf f_x. Given that $S_t = s_t$, the conditional pdf of X_t is $f_x^{[s_t]}$. Given the values of all the state variables, the observation variables are conditionally independent. Therefore, their conditional pdf, given the values of the state variables, is a product pdf,

$$f_{x|s}(\mathbf{x}|\mathbf{s}) = \prod_{t=1}^{N} f_x^{[s_t]}(x_t)$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. To complete the construction, we must obtain a distribution for the state sequence. Such a distribution is confined by only one condition: that the univariate marginal probability mass function (pmf) for each state variable be uniform on the set $\{1, \ldots, M\}$ of states. Assuming that p_s is such a multivariate pmf, we have the pdf of the observations by the law of total probability:

$$f_x^{(N)}(\mathbf{x}) = \sum_{\mathbf{s}} p_s(\mathbf{s}) \prod_{t=1}^{N} f_x^{[s_t]}(x_t) \tag{4.1}$$

where the sum is taken over all M^N possible state sequences. We obtain the cdf by integrating (4.1), to obtain

$$F_x^{(N)}(\mathbf{x}) = \sum_{\mathbf{s}} p_s(\mathbf{s}) \prod_{t=1}^{N} F_x^{[s_t]}(x_t)$$

where $F_x^{[k]}$ is the cdf corresponding to $f_x[k]$.

It is easy to derive the cdf $F_x^{[k]}$. Since $Y^{(k)} \leq x$ if and only if k or more of the values $Y_1, \ldots, Y_M$ are $\leq x$, we have from elementary probability theory that

$$P\{Y^{(k)} \leq x\} = \sum_{j=k}^{M} P\{\text{exactly } j \text{ of the } Y\text{'s are} \leq x\}$$

Since the terms on the right hand side are binomial probabilities with probability of success being $F_x(x)$, we have

$$F_x^{[k]}(x) = \sum_{j=k}^{M} \binom{M}{j} F_x(x)^j [1 - F_x(x)]^{M-j}$$

The pdf $f_x^{[k]}$ is obtained from $F_x^{[k]}$ by differentiation, yielding

$$f_x^{[k]}(x) = f_x(x)\Big\{\sum_{j=k}^{M} j\binom{M}{j} F_x(x)^{j-1}[1 - F_x(x)]^{M-j} - \sum_{j=k}^{M}(M-j)\binom{M}{j} F_x(x)^j[1 - F_x(x)]^{M-j-1}\Big\} \tag{4.2}$$

When computing f_x, one should use the form given in (4.2); the two sums should be computed and then the subtraction performed. This helps to guard against the effect of round-off error, which can cause the values of the pdf to become negative when the true value of the pdf has a very small magnitude.

Equation (4.1) is very general. It is easy to conceive of numerous possibilities for the multivariate pmf p_s. In order for $f_x^{(N)}$ to have f_x as its univariate marginal pdf, it is necessary only that the univariate marginal pmf of p_s be uniform on $\{1, \ldots, M\}$. To simplify our work, however, we shall assume henceforth that the state process is a Markov chain. By definition, the state process is a Markov chain if the conditional distribution of S_t, given the value of S_{t-1} is independent of $\{S_{t-2}, S_{t-3}, \ldots\}$. This implies that the multivariate pmf p_s factors into the product of conditional pmfs,

$$p_s(\mathbf{s}) = \pi_{s_1} \prod_{t=1}^{N-1} p_{s_t s_{t+1}} \tag{4.3}$$

where p_{ij} is the probability that $S_{t+1} = j$ given that $S_t = i$, and π_i is the unconditional probability of S_1 taking the value i. Substituting (4.3) into (4.1), we obtain

$$f_x^{(N)}(\mathbf{x}) = \sum_{s_N=1}^{M} \cdots \sum_{s_1=1}^{M} \pi_{s_1} f_x^{[s_1]}(x_1) \prod_{t=1}^{N-1} \left[p_{s_t s_{t+1}} f_x^{[s_{t+1}]}(x_{t+1})\right] \tag{4.4}$$

The model that we have constructed for the observation process, involving an underlying, unobservable Markov state process and observation process that has a distribution conditional on the state process, is called a *hidden Markov model* (HMM). Such models have been applied to problems in speech processing and other areas [1], [33]. The new twist to

this old concept, however, is the requirement that the unconditional univariate marginal distribution of the observation process be the one prescribed beforehand.

Let $\mathbf{P}$ denote the matrix of transition probabilities. Any Markov chain is completely characterized by its transition probability matrix and the distribution of its initial state. There are certain restrictions on the initial distribution and the transition matrix for our case. An obvious restriction on the matrix $\mathbf{P}$ is that the sum taken along each row be equal to one, that is, $\sum_{j=1}^{M} p_{ij} = 1$. This condition guarantees that the transition probabilities are truly conditional probabilities. Further conditions are required to guarantee that the univariate marginal pmfs are all uniform. First, we require that the initial state be uniformly distributed; thus, $\pi_i = (1/M)$ for all i. Furthermore, we require that the unconditional distribution of each of the succeeding states be uniform. Let $\mathbf{p}_t = (u_{t1}, u_{t2}, \ldots, u_{tM})'$ denote the vector of the unconditional probabilities of S_t, that is, $u_{ti} = P\{S_t = i\}$. According to the definition of a Markov process, these vectors satisfy the recursion

$$\mathbf{P}'\mathbf{p}_{t-1} = \mathbf{p}_t.$$

where $\mathbf{P}'$ is the transpose of $\mathbf{P}$. If $\mathbf{p}_{t-1}$ is uniform, then we have

$$u_{ti} = \sum_{j=1}^{M} p_{ji} u_{t-1,j} = \frac{1}{M} \sum_{j=1}^{M} p_{ji} \tag{4.5}$$

Equation (4.5) reveals that if the column sums of the matrix $\mathbf{P}$ are all equal to 1, $\sum_{j=1}^{M} p_{ji} = 1$, then $\mathbf{p}_t$ is uniform whenever $\mathbf{p}_{t-1}$ is. Therefore, by induction, the unconditional marginal distributions will all be uniform, provided the initial probabilities are uniform and the matrix $\mathbf{P}$ has both row and column sums equal to one. Such a matrix is called doubly stochastic. Some examples of transition matrices are given in a Section sec:examp.

The model just derived is applicable to many situations. In this chapter we refer to it as the first-order HMM/OS, since it is a *H*idden *M*arkov *M*odel derived from *O*rder *S*tatistics. However, in certain situations a more general model may be required. This more general model may be obtained by taking the conditional pdf of X_t given $S_t = s_t$ to be $f_x^{[\nu(s_t)]}$, which is as before except that $\nu(s_t)$ replaces s_t. In this case, the values taken by the state variables are from the set $\{1, 2, \ldots, L\}$ and the function ν takes values in $\{1, 2, \ldots, M\}$. The distribution of the state variables is required to be stationary, but their univariate marginal distribution need not be uniform as long as the distribution of $\nu(S_t)$ is uniform. This model is obviously a generalization, since we may obtain the original model

by taking $L = M$ and ν to be the identity function. The new model allows for higher-order Markov state processes. Hence, we refer to it as a higher-order HMM/OS.

Let $\mathbf{P}$ denote again the matrix of transition probabilities for the higher-order HMM/OS. The requirement that the row sums of the matrix be equal to one holds as before. Let $\mathbf{p}_t = (u_{t1}, u_{t2}, \ldots, u_{tL})$ be the vector of unconditional probabilities for the state S_t. In order for $\nu(S_t)$ to be uniformly distributed, we require that

$$\sum_{j:\nu(j)=k} u_{tj} = \frac{1}{M}, \qquad k = 1, 2, \ldots, M. \tag{4.6}$$

The restrictions on the transition probability matrix, however, are not as clearly defined as they are in the case of the first-order HMM/OS. Instead of the requirement that the column sums be equal to one, we have the requirement that each column satisfy a linear constraint that depends on the initial state distribution. We have some freedom in choosing the initial distribution, being bound only by the constraint (4.6). After the initial distribution is selected, the requirement that the steady-state distribution be equal to the initial distribution places constraints on the transition probabilities. Specifically, we have

$$\pi_i = \sum_{j=1}^{L} p_{ji}\pi_j, \qquad i = 1, 2, \ldots, L$$

where π_j is the probability that the initial state S_1 takes the value j.

4.3 Computation of the HMM/OS Multivariate Densities

The general form of the density given by (4.1) is computationally expensive, since it requires a sum to be taken over M^N possible state sequences. However, when the state process is Markov, the computation may be done very efficiently using a recursive method referred to as the forward-backward algorithm [2], [26], [35], [43]. Using this method, the algorithmic complexity is linear in N, rather than exponential. The forward part of the algorithm is required for computing the value of the density; the backward part and forward part together are required for computing the gradient, which is necessary for some estimation procedures such as maximum likelihood. A variant of the forward-backward algorithm yields the logarithm of the density and its gradient [33].

The algorithm involves the computation of a forward sequence $\alpha_t(j)$ and a backward sequence $\beta_t(j)$. The forward sequence $\alpha_t(j)$, $1 \le t \le N$, $1 \le j \le M$, is defined by the

forward recursion

$$\alpha_1(j) = \pi_j f_x^{[j]}(x_1), \qquad 1 \le j \le M \tag{4.7a}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{M} \alpha_t(i) p_{ij}\right] f_x^{[j]}(x_{t+1}), \qquad 1 \le t \le N-1, 1 \le j \le M \tag{4.7b}$$

The backward sequence $\beta_t(i)$, $1 \le t \le N$, $1 \le i \le M$, is defined by the backward recursion

$$\beta_N(i) = 1, \qquad \text{for } 1 \le i \le M \tag{4.8a}$$

$$\beta_t(i) = \sum_{j=1}^{M} p_{ij} f_x^{[j]}(x_{t+1}) \beta_{t+1}(j), \qquad 1 \le t \le N-1, 1 \le i \le M \tag{4.8b}$$

The pdf (4.4) may be written in terms of the forward and backward sequences as

$$f_x^{(N)}(\mathbf{x}) = \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_t(i) p_{ij} f_x^{[j]}(x_{t+1}) \beta_{t+1}(j) \tag{4.9}$$

for any t such that $1 \le t \le N-1$. Setting $t = N-1$ in (4.9) yields the value of the density in terms of the forward sequence only,

$$f_x^{(N)}(\mathbf{x}) = \sum_{i=1}^{M} \alpha_N(i). \tag{4.10}$$

We can obtain the derivative of $f_x^{(N)}(\mathbf{x})$ with respect to p_{uv} by taking the derivative of (4.10) and (4.7b). Let $\alpha'_t(j)$ denote the derivative of $\alpha_t(j)$ with respect to the parameter p_{uv}. From (4.7b) we have

$$\alpha'_{t+1}(j) = \sum_{i=1}^{M} \alpha'_t(i) p_{ij} f_x^{[j]}(x_{t+1}) + \delta_{jv} \alpha_t(u) f_x^{[v]}(x_{t+1}) \tag{4.11}$$

where $\delta_{jv} = 1$ if $j = v$ and $\delta_{jv} = 0$ if $j \ne v$. Using (4.10) and (4.11) together, we obtain

$$\frac{\partial}{\partial p_{uv}} f_x^{(N)}(\mathbf{x}) = \sum_{j=1}^{M} \alpha'_N(j) \tag{4.12a}$$

$$= \sum_{j=1}^{M} \sum_{i=1}^{M} \alpha'_{N-1}(i) p_{ij} f_x^{[j]}(x_N) + \alpha_{N-1}(u) f_x^{[v]}(x_N) \tag{4.12b}$$

$$= \sum_{i=1}^{M} \alpha'_{N-1}(i) \beta_{N-1}(i) + \alpha_{N-1}(u) f_x^{[v]}(x_N) \tag{4.12c}$$

If we continue to use the recursion (4.11) in (4.12), we eventually obtain

$$\frac{\partial}{\partial p_{uv}} f_x^{(N)}(\mathbf{x}) = \sum_{t=1}^{N-1} \alpha_t(u) f_x^{[v]}(x_{t+1}) \beta_{t+1}(v) \tag{4.13}$$

When N is large, the value of $f_x^{(N)}(\mathbf{x})$ may be such that its value is too small to be represented in many computers. When faced with this problem, one should compute $\log\left[f_x^{(N)}(\mathbf{x})\right]$. The logarithm of the density function, called the log-likelihood function, is also important in its own right, as it is frequently used in maximum likelihood estimation. We now look at a method for computing the log-likelihood ratio directly using the forward recursion sequence $\alpha_t(j)$.

A new sequence $\{\bar{\alpha}_t(j)\}$ is defined by the recursion

$$\bar{\alpha}_1(j) = c_1 \pi_j f_x^{[j]}(x_1), \qquad 1 \le j \le M \tag{4.14a}$$

$$\bar{\alpha}_{t+1}(j) = c_{t+1} \left[\sum_{i=1}^{M} \bar{\alpha}_t(i) p_{ij}\right] f_x^{[j]}(x_{t+1}), \qquad 1 \le t \le N-1, 1 \le j \le M \tag{4.14b}$$

where the sequence $\{c_t\}$ is chosen so that

$$\sum_{j=1}^{M} \bar{\alpha}_t(j) = 1. \tag{4.15}$$

Explicitly, c_t is given by

$$c_t = \left[\sum_{j=1}^{M} \bar{\alpha}_t(j)\right]^{-1}.$$

Comparing (4.14) with (4.7), we see that the sequence $\{\bar{\alpha}_t(j)\}$ is a normalized version of $\{\alpha_t(j)\}$, scaled at each step in the recursion to prevent machine underflow. The explicit relationship between $\bar{\alpha}_t(j)$ and $\alpha_t(j)$ is

$$\bar{\alpha}_t(j) = c_1 c_2 \cdots c_t \alpha_t(j). \tag{4.16}$$

This result is proved by induction. Clearly, $\bar{\alpha}_1(j) = c_1 \alpha_1(j)$. If equation (4.16) holds for a particular value of t, then for $t+1$, we have

$$\bar{\alpha}_{t+1}(j) = c_{t+1} \left[\sum_{i=1}^{M} \bar{\alpha}_t(i) p_{ij}\right] f_x^{[j]}(x_{t+1}) \tag{4.17a}$$

$$= c_1 c_2 \cdots c_{t+1} \left[\sum_{i=1}^{M} \alpha_t(i) p_{ij}\right] f_x^{[j]}(x_{t+1}) \tag{4.17b}$$

$$= c_1 c_2 \cdots c_{t+1} \alpha_{t+1}(j) \tag{4.17c}$$

Now from (4.10), (4.16), and (4.15), we have

$$f_x^{(N)}(\mathbf{x}) = \sum_{i=1}^{M} \alpha_N(i) = (c_1 c_2 \cdots c_N)^{-1} \sum_{i=1}^{M} \bar{\alpha}_N(i) = (c_1 c_2 \cdots c_N)^{-1} \tag{4.18}$$

Finally, we have

$$\log\left[f_x^{(N)}(\mathbf{x})\right] = -\sum_{t=1}^{T} \log\, c_t \tag{4.19}$$

Thus to compute the log-likelihood function, rather than the density function, one must compute the scale factor c_t and normalize $\alpha_t(j)$, $j = 1, \ldots, M$, at each step. The sum in (4.19) may be accumulated as the recursion proceeds.

We may also use the scaled forward sequence $\{\bar{\alpha}_t(j)\}$ together with a scaled backward sequence $\{\bar{\beta}_t(i)\}$ to compute the derivative of the log-likelihood function. The scaled backward sequence uses the same scale factors $\{c_t\}$ as the forward sequence, and has the recursive definition

$$\bar{\beta}_N(i) = c_N, \qquad 1 \le i \le M \tag{4.20a}$$

$$\bar{\beta}_t(i) = c_t \sum_{j=1}^{M} p_{ij} f_x^{[j]}(x_{t+1}) \bar{\beta}_{t+1}(j), \qquad 1 \le t \le N-1, 1 \le i \le M \tag{4.20b}$$

The explicit relationship between $\bar{\beta}_t(i)$ and $\beta_t(i)$ is

$$\bar{\beta}_t(i) = c_t c_{t+1} \cdots c_N \beta_t(i), \tag{4.21}$$

which may be proved by induction in the same way (4.16) was proved. Using (4.16), (4.21), and (4.13), we may write

$$\frac{\partial}{\partial p_{uv}} f_x^{(N)}(\mathbf{x}) = (c_1 c_2 \cdots c_N)^{-1} \sum_{t=1}^{N-1} \bar{\alpha}_t(u) f_x^{[v]}(x_{t+1}) \bar{\beta}_{t+1}(v) \tag{4.22}$$

Since

$$\frac{\partial}{\partial p_{uv}} \log f_x^{(N)}(\mathbf{x}) = \frac{\frac{\partial}{\partial p_{uv}} f_x^{(N)}(\mathbf{x})}{f_x^{(N)}(\mathbf{x})},$$

we may combine (4.18) and (4.22) to obtain

$$\frac{\partial}{\partial p_{uv}} \log f_x^{(N)}(\mathbf{x}) = \sum_{t=1}^{N-1} \bar{\alpha}_t(u) f_x^{[v]}(x_{t+1}) \bar{\beta}_{t+1}(v).$$

The forward-backward algorithm is easily adapted to compute the density and gradient of the higher-order HMM/OS. Equation (4.7) is changed to become

$$\alpha_1(j) = \pi_j f_x^{[\nu(j)]}(x_1), \qquad \text{for } 1 \le j \le M \tag{4.23a}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{M} \alpha_t(i) p_{ij}\right] f_x^{[\nu(j)]}(x_{t+1}), \qquad \text{for } 1 \le t \le N-1,\ 1 \le j \le M \tag{4.23b}$$

and (4.8) is changed to become

$$\beta_N(i) = 1, \qquad \text{for } 1 \le i \le M \tag{4.24a}$$

$$\beta_t(i) = \sum_{j=1}^{M} p_{ij} f_x^{[\nu(j)]}(x_{t+1}) \beta_{t+1}(j), \qquad \text{for } 1 \le t \le N-1,\ 1 \le i \le M \tag{4.24b}$$

The other equations are changed in a similar way by replacing $f_x^{[j]}$ by $f_x^{[\nu(j)]}$.

Finally, we consider efficient ways for generating the observations. Two methods are possible. Common to both methods is task of generating the state process. When the current state S_t has the value i, then the ith row of the transition probability matrix contains the pmf of the next state. The task of generating the state process, then, involves generating a sequence of discrete random variates according to the probabilities in the transition probability matrix. To generate the observations, we may generate a set of M independent random variates from the density f_x, sort them to obtain the order statistics, and choose the order statistic selected by the current state as the current observation. This is a straightforward imitation of the mechanism by which our HMM was constructed in Section 4.2. One commonly used method for generating random variates is to generate a uniform variate U and then obtain the desired variate Y by evaluating the inverse cdf, $Y = F^{-1}(U)$. If such a method is to be used then the inverse cdf need not be applied until the final order statistic has been selected. An alternative method for generating the observations, which may be more efficient when M is large, is to generate the current state s_t and then the current observation directly from the distribution of the s_t-th order statistic. Again, if the inverse cdf method is to be used to generate the observation variate, then the inverse cdf should be applied at the final step. Thus we generate U from the distribution of the s_t-th order statistic and then the observation $X_t = F_x^{-1}(U)$.

4.4 Examples

The number of parameters in an HMM/OS distribution can be very large. In some situations where we require a very close fit of the model to data, we may prefer a large number of parameters. However, in other situations it may be more desirable to have a 'parsimonious' model, that is, one that has just a handful of parameters. One advantage of a parsimonious model is that estimation of the parameters is greatly simplified. We now look at several examples of parsimonious models.

Our first example is one that requires only two parameters. These parameters are M and p, where M is the number of states, and p is a parameter that defines the transition probability matrix. If we take $M = 4$ and $q = 1 - 2p$, then the transition probability matrix is

$$\mathbf{P} = \begin{bmatrix} p+q & q & 0 & 0 \\ q & p & q & 0 \\ 0 & q & p & q \\ 0 & 0 & q & p+q \end{bmatrix} \tag{4.25}$$

This transition probability matrix yields a positively correlated bivariate distribution for (X_t, X_{t+1}). We could, alternatively, define a negatively correlated bivariate distribution by rotating the matrix in (4.25) by 90 degrees. This yields

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & q & p+q \\ 0 & q & p & q \\ q & p & q & 0 \\ p+q & q & 0 & 0 \end{bmatrix}$$

Both the parameters p and M determine the strength of the correlation. For a fixed M, the correlation increases with increasing p, while for a fixed p, the correlation increases with increasing M.

Numerous variations of the two parameter model are possible. First, we note that any permutation of the rows or columns yields a legitimate transition probability matrix. Second, we could define $q = (1-p)/(M-1)$ and the transition probability matrix as

$$\mathbf{P} = \begin{bmatrix} p+q & q & q & q \\ q & p & q & q \\ q & q & p & q \\ q & q & q & p+q \end{bmatrix}$$

Other possibilities arise as subcases of three parameter models by taking two of the parameters equal.

A three parameter model requires the parameters M, p, and q, where M is the number of states and p and q define the transition probability matrix. For a specific example, let $M = 5$ and $r = (1 - p - 2q)/2$. Then the transition probability matrix is

$$\mathbf{P} = \begin{bmatrix} p+q+r & q & r & 0 & 0 \\ q & p+r & q & r & 0 \\ r & q & p & q & r \\ 0 & r & q & p+r & q \\ 0 & 0 & r & q & p+q+r \end{bmatrix} \tag{4.26}$$

This example yields a positive correlation between X_t and X_{t+1}. Rotating the matrix in (4.26) by 90 degrees yields a negative correlation. One simple variation of (4.26) is to replace all the 0's by r's. This yields

$$\mathbf{P} = \begin{bmatrix} p+q+r & q & r & r & r \\ q & p+r & q & r & r \\ r & q & p & q & r \\ r & r & q & p+r & q \\ r & r & r & q & p+q+r \end{bmatrix}$$

and we require that $r = (1 - 2q)/(M - 3)$. Other variations involve the transition probabilities for the extreme states. For example, instead of (4.26), we could have

$$\mathbf{P} = \begin{bmatrix} p+q & q+r & r & 0 & 0 \\ q+r & p & q & r & 0 \\ r & q & p & q & r \\ 0 & r & q & p & q+r \\ 0 & 0 & r & q+r & p+q \end{bmatrix}$$

Parsimonious models with four or more parameters are easily constructed using the ideas above. The basic idea is to keep the diagonals as constant as possible and find a way at the ends of the diagonals to accommodate the constraints on the row and column sums. While the examples above are symmetric, symmetry is not necessary.

The next example is a non-symmetric model. Let $M = 4$ and let $\mathbf{P}$ be defined as

$$\mathbf{P} = \begin{bmatrix} p & r & r & q \\ q & p & r & r \\ r & q & p & r \\ r & r & q & p \end{bmatrix} \tag{4.27}$$

where p and q are the independently chosen parameters and $r = (1 - p - q)/(M - 2)$. If r is small relative to p and q, the states will tend to decrease down to state 1 and then jump up to state M.

4.5 Parameter Estimation

There is little reason to suspect that the HMMs derived in this chapter would not work well as models for discrimination. That the models are sufficiently flexible seems evident. However, flexibility is useless unless the models can be well fitted to the data. Therefore, fitting the models to data, or estimating the model parameters, is perhaps the most important task. In this section, we consider three different approaches to parameter estimation: least squares, minimum discrimination information, and maximum likelihood.

Our main focus is on estimating the entries of the transition probabilty matrix. In practice, we must also deal with the number M of order statistics and the parameters of the univariate marginal distribution. The task of identifying M may be viewed as a task of model selection analogous to the task of selecting the model orders for ARMA models. We do not consider this task here. We do, however, compare simulation results for different values of M. The task of estimating the parameters of the univariate marginal distribution is also a task that we do not address.

4.5.1 Least Squares Estimation

Least squares estimation represents a straightforward approach to the estimation problem. Friedlander and Porat [13], as well as other researchers, have used the method to estimate the parameters of ARMA models. We develop here the least squares method for estimating the parameters for a first-order HMM/OS. The method we derive has the advantages of being simple, computationally inexpensive, and in certain cases very robust. The idea is similar to that of the classical method of moments, which turns out to be a

subcase of our least squares method. The discussion given here involves only the bivariate pdf for a first-order HMM/OS,

$$f_x^{(2)}(x,y) = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M} p_{ij} f_x^{[i]}(x) f_x^{[j]}(y) \tag{4.28}$$

where p_{ij} are the parameters to be estimated. The parameters of the bivariate pdf are sufficient to determine the multivariate pdf for a first-order model.

Let $\{(x_i, y_i), i = 1, 2, \ldots, n\}$ denote the sample data we have available for estimation. The data pairs may be independent of each other; however, this assumption is not required, and the pairs may be obtained from a single time series, so that, for example, $x_{i+1} = y_i$. Let g_k, $k = 1, 2, \ldots, K$, be a collection of real-valued functions on $\mathbb{R}^2$ which we have chosen subjectively and which we refer to generically as *moment functions*. For example, each g_k could have the form $g_k(x,y) = x^l y^m$ for some nonnegative integers l and m. The statistics we use for estimation are the sample moments

$$\hat{r}_k = \frac{1}{n}\sum_{i=1}^{n} g_k(x_i, y_i) \qquad k = 1, 2, \ldots, K$$

computed from the collected data. The sample moments $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_K)$ are approximations to the true moments $\mathbf{r} = (r_1, \ldots, r_K)$ of the unknown pdf, defined by

$$r_k = \mathrm{E}\, g_k(X,Y) = \iint g_k(x,y) f_x^{(2)}(x,y)\, dx\, dy \qquad k = 1, 2, \ldots, K$$

where (X, Y) have the bivariate pdf (4.28).

The moment vector $\mathbf{r}$ is a linear function of the parameter vector $\mathbf{p} = (p_{11}, p_{12}, \ldots, p_{MM})$; thus

$$\mathbf{C}\mathbf{p} = \mathbf{r} \tag{4.29}$$

where $\mathbf{C}$ is a matrix with elements

$$\mathbf{C}_{ij} = \frac{1}{M}\iint_{\mathbb{R}^2} g_i(x,y) f_x^{[\alpha]}(x) f_x^{[\beta]}(y)\, dx\, dy \tag{4.30}$$

that we assume to have full column rank. Seemingly, the simplest estimation method would be to compute $\hat{\mathbf{r}}$, substitute it for $\mathbf{r}$ in (4.29), and solve for $\mathbf{p}$. In practice, though, this does not always work because the resulting solution $\mathbf{p}$ may violate certain required constraints. These constraints include the linear equality constraints, $\sum_i p_{ij} = 1$, $j = 1, \ldots, M$ and $\sum_j p_{ij} = 1$, $i = 1, \ldots, M$, and the simple bounds $p_{ij} \geq 0$. Even if the linear equality

constraints are included in the system of linear equations that are solved to obtain $\mathbf{p}$, there is no guarantee that all the components of $\mathbf{p}$ will be nonnegative.

As a straighforward modification to prevent violation of the constraints, we recast the problem as a linearly constrained linear least squares problem. Although we cannot always solve $\mathbf{Cp} = \hat{\mathbf{r}}$ to obtain an acceptible $\mathbf{p}$, we can always find $\mathbf{p}$ to minimize $\|\mathbf{Cp} - \hat{\mathbf{r}}\|^2$, where $\|\cdot\|$ is the Euclidean norm, subject to the constraints on $\mathbf{p}$. Writing $\|\mathbf{Cp} - \hat{\mathbf{r}}\|^2$ as a quadratic function, the new problem, formally stated, is

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\mathbf{p}'\mathbf{H}\mathbf{p} + \mathbf{g}'\mathbf{p} + c & \\ \text{subject to} & \sum_{i=1}^{M} p_{ij} = 1 & 1 \le i \le M \\ & \sum_{j=1}^{M} p_{ij} = 1 & 1 \le j \le M-1 \\ & p_{ij} \ge 0 & 1 \le i,j \le M \end{array} \tag{4.31}$$

where $\mathbf{H} = 2\mathbf{C}'\mathbf{C}$, $g = -2\mathbf{C}'\hat{\mathbf{r}}$, and $c = \hat{\mathbf{r}}'\hat{\mathbf{r}}$. The matrix $\mathbf{H}$ is positive semi-definite, since $\mathbf{x}'\mathbf{H}\mathbf{x} = \|\mathbf{Cx}\|^2 \ge 0$, and it is positive definite if and only if $\mathbf{C}$ has full column rank (or, equivalently, the null space of $\mathbf{C}$ contains no nonzero elements). That $\mathbf{H}$ is positive definite is important, since this means that we may use any one of a number of available efficient algorithms [16] for solving quadratic programming problems. Note that one of the equality constraints in (4.31) has been eliminated, since it is implied by the other constraints.

A rather obvious fact that deserves some discussion is the over-parameterization of the HMM pdf. Because of the linear constraints on the transition probabilities, some of them are redundant, since they may be expressed as a linear function of the other transition probabilities. The over-parameterization is not necessarily a situation to be avoided in practice, because some optimization algorithms operate more efficiently with the over-parameterized representation than with the more complex reduced parameter representation. Nevertheless, the reduced parameter representation is important to the discussion that follows. Furthermore, the parsimonious models that were given in the preceding section are essentially reduced parameter representations, and therefore, by introducing a general form for the reduced parameter representations, they require no special treatment.

Our reduced parameter representation is based on the orthogonal decomposition of $\mathbb{R}^{M^2}$ with respect to the null space of the constraint coefficient matrix. Let us denote the system of linear equality constraints by

$$\mathbf{Ap} = \mathbf{b} \tag{4.32}$$

where $\mathbf{p} = (p_{11}, p_{12}, \ldots, p_{MM})$ is the vector of transition probabilities. Assume, without loss of generality, that the rows of $\mathbf{A}$ are linearly independent. This assumption simply means that there are no redundant constraints. According to our assumptions, the linear equality constraints contain, as a minimum, the constraints that the row and column sums of $\mathbf{P}$ be equal to 1. These constraints are linearly dependent, so that one of them must be eliminated. Thus, we have at least $2M - 1$ linear equality constraints. Other constraints are included in the case of a parsimonious model. Let the dimensions of $\mathbf{A}$ be $n \times m$, where $m = M^2$ and $n < m$. Of the m parameters in $\mathbf{p}$, n of them are redundant and may be eliminated. The remaining $d = m - n$ parameters are free. The vector $\mathbf{P}$ may be represented

$$\mathbf{p} = \mathbf{Y}\mathbf{p}_y + \mathbf{Z}\mathbf{p}_z \tag{4.33}$$

where $\mathbf{Z}$ is an $m \times d$ matrix with orthogonal columns that span the null space of $\mathbf{A}$ and $\mathbf{Y}$ is an $m \times n$ matrix with orthogonal columns that are orthogonal to the columns of $\mathbf{Z}$. Such a decomposition is usually obtained in practice by the QR algorithm [16]. It is a basic result from linear algebra that the columns of $\mathbf{Y}$ span the range space of $\mathbf{A}'$. Quite obviously $\mathbf{Z}\mathbf{p}_z$ lies in the null space of $\mathbf{A}$, while $\mathbf{Y}\mathbf{p}_y$ lies in the range space of $\mathbf{A}'$. The vector $\mathbf{p}_y$ remains the same for any $\mathbf{p}$ that satisfies (4.32) and is determined by solving

$$\mathbf{A}\mathbf{Y}\mathbf{p}_y = \mathbf{b}. \tag{4.34}$$

Equation (4.34) is obtained from (4.32), (4.33), and the fact that the columns of $\mathbf{Z}$ are orthogonal to the rows of $\mathbf{A}$. On the other hand, $\mathbf{p}_z$ represents the independent parameters in our re-parameterization. Because $\mathbf{Z}$ has orthogonal columns, $\mathbf{Z}'\mathbf{Z} = I$, so that we may obtain $\mathbf{p}_z$ from $\mathbf{p}$ by

$$\mathbf{p}_z = \mathbf{Z}'(\mathbf{p} - \mathbf{Y}\mathbf{p}_y)$$

and we may obtain $\mathbf{p}$ from $\mathbf{p}_z$ by (4.33).

In some situations, it may be worthwhile to further generalize the idea of linear least squares to that of weighted linear least squares. The idea is to choose a symmetric positive definite weight matrix W that assigns different weights to the different residuals. While the standard least squares method minimizes $\|\mathbf{C}\mathbf{p} - \hat{\mathbf{r}}\|^2 = (\mathbf{C}\mathbf{p} - \hat{\mathbf{r}})'(\mathbf{C}\mathbf{p} - \hat{\mathbf{r}})$, the weighted least squares method minimizes $(\mathbf{C}\mathbf{p} - \hat{\mathbf{r}})'W(\mathbf{C}\mathbf{p} - \hat{\mathbf{r}})$. The two methods are the same if W is chosen as the identity matrix. Freidlander and Porat [12], [13] discuss the optimal selection of the weight matrix W for unconstrained least squares estimation. The

matrix they give is optimal in that the resulting estimator has the smallest asymptotic variance. In the remainder of this subsection we adapt their result to our problem.

Let $\theta = \mathbf{p}_z$ denote the d-dimensional reduced vector of parameters to be estimated. The weighted least squares estimation requires that we minimize $(\mathbf{Cp} - \hat{\mathbf{r}})'\mathbf{W}(\mathbf{Cp} - \hat{\mathbf{r}})$ as a function of $\mathbf{p}$ subject to certain linear equality constraints and bounds on the variables. Written in terms of θ, the cost function to be minimized is $(\mathbf{CZ}\theta + \mathbf{CY}\mathbf{p}_y - \hat{\mathbf{r}})'\mathbf{W}(\mathbf{CZ}\theta + \mathbf{CY}\mathbf{p}_y - \hat{\mathbf{r}})$ according to (4.33). Because the redundant parameters in $\mathbf{p}$ have been eliminated, the minimization of the new cost function is subject to linear inequality constraints only. These constraints require that each component of $\mathbf{p} = \mathbf{Y}\mathbf{p}_y + \mathbf{Z}\theta$ be nonnegative. In the reduced representation, the constraints are $\mathbf{Z}\theta \geq -\mathbf{Y}\mathbf{p}_y$, where the inequality is applied to the individual components of the vectors. Thus, the weighted least squares estimate $\hat{\theta}$ is obtained as the solution of the problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\theta'\mathbf{H}\theta + \mathbf{g}'\theta + c \\ \text{subject to} & \mathbf{Z}\theta \geq -\mathbf{Y}\mathbf{p}_y \end{array} \tag{4.35}$$

where $\mathbf{H} = 2\mathbf{Z}'\mathbf{C}'\mathbf{W}\mathbf{C}\mathbf{Z}$, $\mathbf{g} = 2\mathbf{Z}'\mathbf{C}'\mathbf{W}(\mathbf{C}\mathbf{Y}\mathbf{p}_y - \hat{\mathbf{r}})$, and $c = (\mathbf{C}\mathbf{Y}\mathbf{p}_y - \hat{\mathbf{r}})'\mathbf{W}(\mathbf{C}\mathbf{Y}\mathbf{p}_y - \hat{\mathbf{r}})$. Since $\mathbf{y}'\mathbf{H}\mathbf{y} = 2\|\mathbf{W}^{1/2}\mathbf{C}\mathbf{Z}\mathbf{y}\|^2$, $\mathbf{H}$ is strictly positive definite if and only if $\mathbf{C}$ has full column rank.

When considering any estimation method, the question arises as to whether the estimate is consistent, that is, whether $\hat{\theta} \to \theta$ as $n \to \infty$. If the data are i.i.d. and each moment function g_k has finite second moment, $\mathrm{E}\,|g_k(X,Y)|^2 < \infty$, then by the Strong Law of Large Numbers the sample moments converge to the true moments almost surely, $\hat{\mathbf{r}} \to \mathbf{r}$, as $n \to \infty$. To prove the consistency of the estimate then requires that we show that the mapping $\hat{\mathbf{r}} \to \hat{\theta}$ is continuous. The following theorem states that the estimate obtained by solving the minization problem (4.35) is consistent.

Theorem 4.2 *Let* $\mathbf{C}$*, with elements given by (4.30), have full column rank, and suppose*

$$\iint |g_k(x,y)|^2 f_x^{(2)}(x,y)\,dx\,dy < \infty$$

for $k = 1,\ldots,K$*. Then as* $n \to \infty$*,* $\hat{\theta}_n \to \theta$ *almost surely.*

Proof. Under the hypothesis that $\mathrm{E}|g_k(x,y)|^2 < \infty$ for $k = 1,\ldots,K$, the strong law of large numbers implies that $\hat{\mathbf{r}} \to \mathbf{r}$ a.s. Since $\theta = \hat{\theta}(\mathbf{r})$, the question to be resolved is whether $\hat{\theta}(\hat{\mathbf{r}})$, as a function of $\hat{\mathbf{r}}$, is continuous. Let us denote the cost function of (4.35) by $F(\theta, \hat{\mathbf{r}})$.

Because $\mathbf{C}$ has full column rank, F is a positive definite quadratic function. For a given $\hat{\mathbf{r}}$, let θ^* be the vector that minimizes $F(\cdot, \hat{\mathbf{r}})$ without regard for the inequality constaints. By setting the derivative of F equal to zero, we obtain θ^* explicity as $\theta^* = -\mathbf{H}^{-1}\mathbf{g}$, where $\mathbf{g}$, as we recall, is a linear function of $\hat{\mathbf{r}}$. Therefore, θ^* is a continuous function of $\hat{\mathbf{r}}$ and $\theta^* \to \theta$ a.s.

From a Taylor series expansion about θ^*, we obtain $F(\mathbf{y}, \hat{\mathbf{r}}) - F(\theta^*, \hat{\mathbf{r}}) = \frac{1}{2}(\mathbf{y} - \theta^*)'\mathbf{H}(\mathbf{y} - \theta^*)$, and since $\mathbf{H}$ is strictly positive definite, we have

$$\lambda_1 \leq \frac{F(\mathbf{y}, \hat{\mathbf{r}}) - F(\theta^*, \hat{\mathbf{r}})}{\|\mathbf{y} - \theta^*\|^2} \leq \lambda_2 \tag{4.36}$$

for some $0 < \lambda_1 \leq \lambda_2 < \infty$.

Since $\hat{\theta}$ minimizes $F(\cdot, \hat{\mathbf{r}})$ subject to the linear inequality constraints, and since θ satisfies these constraints, we have $F(\hat{\theta}, \hat{\mathbf{r}}) \leq F(\theta, \hat{\mathbf{r}})$. Therefore, we have

$$\lambda_2\|\theta - \theta^*\|^2 \geq F(\theta, \hat{\mathbf{r}}) - F(\theta^*, \hat{\mathbf{r}}) \geq F(\hat{\theta}, \hat{\mathbf{r}}) - F(\theta^*, \hat{\mathbf{r}}) \geq \lambda_1\|\hat{\theta} - \theta^*\|^2 \tag{4.37}$$

And finally,

$$\|\hat{\theta} - \theta\| \leq \|\theta^* - \theta\| + \|\hat{\theta} - \theta^*\| \leq \left(1 + \sqrt{\lambda_2/\lambda_1}\right)\|\theta - \theta^*\| \to 0 \tag{4.38}$$

Q.E.D.

The theorem states that the estimate $\hat{\theta}$ is consistent for any positive definite weight matrix $\mathbf{W}$. The following results concern the optimal choice of the weight matrix $\mathbf{W}$ and follow in three stages. First, we obtain an expression for the asymptotic variance of the estimate $\hat{\theta}$, an expression that explicitly involves $\mathbf{W}$. This result was presented by Friedlander and Porat in [12], but their proof is in error. A correct and somewhat more involved proof is given here for our special case. Second, we present a lower bound on the asymptotic variance that is independent of $\mathbf{W}$. And finally, we present a weight matrix that attains the lower bound on the asymptotic variance. These last two results are adapted from [12].

Before moving on, it is a good idea to review old notation, introduce some new notation, and state explicitly certain relationships among some of the variables of interest. This section starts with the basic assumption that the distribution of the observed data is determined by the unknown parameter $\theta \in \Theta$. The moment vector $\mathbf{r}$ is related to θ by the linear function

$$\mathbf{r} = f(\theta) \stackrel{\text{def}}{=} \mathbf{C}\mathbf{Z}\theta + \mathbf{C}\mathbf{Y}\mathbf{p}_y \tag{4.39}$$

where $\mathbf{p}_y$ is chosen so that the vector $\mathbf{p} = \mathbf{Z}\theta + \mathbf{Y}\mathbf{p}_y$ satisfies the linear equality constraints. The requirement that the components of $\mathbf{p}$ be nonnegative defines a convex set Θ of allowable values for θ.

According to our estimation method, we obtain an estimate $\hat{\mathbf{r}}$ of $\mathbf{r}$, and from $\hat{\mathbf{r}}$ we obtain an estimate $\hat{\theta}_n$ of θ by an inverse relationship to (4.39). In the interesting case, and in the case we consider here, the dimension of $\mathbf{r}$ is greater than the dimension of θ. Consequently, there is no unique inverse to (4.39). Any inverse h that we care to define should agree with (4.39), so that $h\big[f(\theta)\big] = \theta$ for all $\theta \in \Theta$; however, outside the set $f(\Theta) = \{\mathbf{r} : \mathbf{r} = f(\theta), \theta \in \Theta\}$ the definition of h is not unique. The minimization problem (4.35), for example, expresses such an inverse function $\hat{\theta}_n = h(\hat{\mathbf{r}}_n)$ that is compatible with f and is therefore uniquely defined on $f(\Theta)$. Outside $f(\Theta)$, however, h depends on the weight matrix $\mathbf{W}$.

The non-uniqueness of the inverse function to f suggests that we look for an "optimal" weight matrix $\mathbf{W}$. The criterion for choosing the optimal is the asymptotic variance of the estimate. We now present a theorem that states an explicit expression for the asymptotic variance of the estimate $\hat{\theta}_n$. The asymptotic variance of the estimate θ_n^*, which, as we defined earlier, is the solution of the least squares problem without regard for the constraints, is straightforward to obtain. The idea involved in obtaining the asymptotic variance for $\hat{\theta}_n$ is this: when θ is in the interior Θ° of Θ, the boundary effect that distinguishes $\hat{\theta}_n$ from θ_n^* becomes negligible as $n \to \infty$. Thus $\hat{\theta}_n$ and θ_n^* have the same asymptotic variance. The formal proof of this result requires the following lemmas.

Lemma 4.3 *Suppose that the fourth-order central moments of θ_n^* are $o(n)$ and that $\theta \in \Theta^\circ$. Let $\eta > 0$ and let A_n be the event $\{\|\theta_n^* - \theta\| \geq \eta\}$. Then $\lim_{n\to\infty} nPA_n = 0$.*

Proof. Using the Markov inequality, we have

$$\begin{aligned} nPA_n &\leq n\eta^{-4}\mathrm{E}\,\|\theta_n^* - \theta\|^4 \\ &= n\eta^{-4}\sum_{i=1}^{d}\sum_{j=1}^{d}\mathrm{E}\,(\theta_{ni}^* - \theta_i)^2(\theta_{nj}^* - \theta_j)^2 \end{aligned} \tag{4.40}$$

Under the assumption that the fourth-order central moments of θ_n^* are $o(n)$, the last expression in (4.40) converges to 0. Q.E.D.

Lemma 4.4 *There exists a $K > 0$ such that $\|\hat{\theta}_n - \theta\| \leq K\|\theta_n^* - \theta\|$.*

Proof. As in the proof of Theorem 4.2 let us again denote the cost function of (4.35) by $F(\theta, \hat{\mathbf{r}})$. Then, as before, we have

$$\lambda_1 \leq \frac{F(\mathbf{y}, \hat{\mathbf{r}}) - F(\theta^*, \hat{\mathbf{r}})}{\|\mathbf{y} - \theta^*\|^2} \leq \lambda_2$$

for some $0 < \lambda_1 \leq \lambda_2 < \infty$. Following again the steps of (4.37) and (4.38), we have the desired result with $K = (1 + \sqrt{\lambda_2/\lambda_1})$. Q.E.D.

Lemma 4.5 *Suppose that the fourth-order central moments of θ_n^* are $o(n)$ and that $\theta \in \Theta^\circ$. Then the asymptotic covariance matrices of $\hat{\theta}_n$ and θ_n^* are equal, that is, for all k and l*

$$\lim_{n\to\infty} n\mathrm{E}\,(\hat{\theta}_{nk} - \theta_k)(\hat{\theta}_{nl} - \theta_l) = \lim_{n\to\infty} n\mathrm{E}\,(\theta_{nk}^* - \theta_k)(\theta_{nl}^* - \theta_l)$$

Proof. Let A_n be as in Lemma 4.3. Then we have

$$\begin{aligned} n\Big|\mathrm{E}\,(\hat{\theta}_{nk} - \theta_k)(\hat{\theta}_{nl} - \theta_l) - (\theta_{nk}^* - \theta_k)(\theta_{nl}^* - \theta_l)\Big| \\ \leq n\Big|\mathrm{E}\, I_{A_n^c}\Big[(\hat{\theta}_{nk} - \theta_k)(\hat{\theta}_{nl} - \theta_l) - (\theta_{nk}^* - \theta_k)(\theta_{nl}^* - \theta_l)\Big]\Big| \\ + n\Big|\mathrm{E}\, I_{A_n}(\hat{\theta}_{nk} - \theta_k)(\hat{\theta}_{nl} - \theta_l)\Big| \\ + n\Big|\mathrm{E}\, I_{A_n}(\theta_{nk}^* - \theta_k)(\theta_{nl}^* - \theta_l)\Big| \end{aligned} \tag{4.41}$$

If $\theta_n^* \in \Theta^\circ$, then $\hat{\theta}_n = \theta_n^*$. Thus if η is sufficiently small, the first term on the right-hand side of (4.41) is zero for all n. We apply the Cauchy-Schwartz inequality to the third term to obtain

$$n^2\Big|\mathrm{E}\, I_{A_n}(\theta_{nk}^* - \theta_k)(\theta_{nl}^* - \theta_l)\Big|^2 \leq \Big[nPA_n\Big]\Big[n\mathrm{E}\,(\theta_{nk}^* - \theta_k)^2(\theta_{nl}^* - \theta_l)^2\Big]$$

The first factor on the right-hand side above converges to zero by Lemma 4.3, while the second factor converges to zero by the hypothesis that the fourth-order moments of θ_n^* are $o(n)$.

Finally, we consider the second term in (4.41). Applying the Cauchy-Schwartz inequality and the result of Lemma 4.4, we have

$$\begin{aligned} n^2\Big|\mathrm{E}\, I_{A_n}(\hat{\theta}_{nk} - \theta_k)(\hat{\theta}_{nl} - \theta_l)\Big|^2 &\leq \Big[nPA_n\Big]\Big[n\mathrm{E}\,(\hat{\theta}_{nk} - \theta_k)^2(\hat{\theta}_{nl} - \theta_l)^2\Big] \\ &\leq \Big[nPA_n\Big]\Big[n\mathrm{E}\,\|\hat{\theta}_n - \theta\|^4\Big] \\ &\leq K^4\Big[nPA_n\Big]\Big[n\mathrm{E}\,\|\theta_n^* - \theta\|^4\Big] \end{aligned}$$

Since $\mathrm{E}\,\|\theta_n^* - \theta\|^4$ is a linear combination of fourth-order moments of θ_n^*, the second term of the right-most expression above converges to zero. Q.E.D.

Finally, we consider conditions that guarantee that the central moments of θ_n^* are $o(n)$, as required by the preceding lemmas.

Lemma 4.6 *Suppose that* $\mathrm{E}\,g_k(X_1, X_2)^4 \leq \Delta < \infty$ *for* $k = 1, 2, \ldots, K$ *and that the underlying state process is ergodic. Then the fourth-order central moments of* θ_n^* *are* $o(n)$.

Proof. Let $\bar{g}_k(x, y) = g_k(x, y) - \mathrm{E}\,g_k(X_1, X_2)$ be a centered version of g_k. From the hypothesis that $\mathrm{E}\,g_k(X_1, X_2)^4 \leq \Delta < \infty$ for $k = 1, 2, \ldots, K$, it follows that $\mathrm{E}\,\bar{g}_k(X_1, X_2)^4 \leq \Delta' < \infty$ for $k = 1, 2, \ldots, K$. First, we show that

$$\left|\mathrm{E}\,\bar{g}_\alpha(X_i, X_{i+1})\,\bar{g}_\beta(X_j, X_{j+1})\,\bar{g}_\gamma(X_k, X_{k+1})\,\bar{g}_\delta(X_l, X_{l+1})\right| \leq \Delta'. \tag{4.42}$$

Simplifying our notation, we have by the Cauch-Schwartz inequality that

$$\begin{aligned}\left|\mathrm{E}\,\bar{g}_\alpha\,\bar{g}_\beta\,\bar{g}_\gamma\,\bar{g}_\delta\right|^4 &\leq \left|\mathrm{E}\,\bar{g}_\alpha^2\,\bar{g}_\beta^2\right|^2 \left|\mathrm{E}\,\bar{g}_\gamma^2\,\bar{g}_\delta^2\right|^2 \\ &\leq \mathrm{E}\,\bar{g}_\alpha^4\,\mathrm{E}\,\bar{g}_\beta^4\,\mathrm{E}\,\bar{g}_\gamma^4\,\mathrm{E}\,\bar{g}_\delta^4 \\ &\leq (\Delta')^4\end{aligned} \tag{4.43}$$

which proves (4.42). Similarly, the moments of $\bar{g}_\alpha$, $\bar{g}_\beta$, and $\bar{g}_\gamma$ of orders two and three are also bounded. Next, we show that for $i < j < k < l - q$, we have

$$\left|\mathrm{E}\,\bar{g}_\alpha(X_i, X_{i+1})\,\bar{g}_\beta(X_j, X_{j+1})\,\bar{g}_\gamma(X_k, X_{k+1})\,\bar{g}_\delta(X_l, X_{l+1})\right| \leq K' e^{-cq}. \tag{4.44}$$

for some constants K' and c. We show this by conditioning on the underlying state process. Thus, with $p_{uv}^{(\nu)}$ denoting the (u, v)th entry in the ν-step transition probability matrix, we have

$$\begin{aligned}\left|\mathrm{E}\,\bar{g}_\alpha\,\bar{g}_\beta\,\bar{g}_\gamma\,\bar{g}_\delta\right| &= \left|\mathrm{E}\,\bar{g}_\alpha\,\bar{g}_\beta\,\bar{g}_\gamma\,\bar{g}_\delta - \mathrm{E}\,\bar{g}_\alpha\,\bar{g}_\beta\,\bar{g}_\gamma\,\mathrm{E}\,\bar{g}_\delta\right| \\ &= \left|\frac{1}{M}\sum_{u=1}^{M}\sum_{v=1}^{M}\mathrm{E}[\bar{g}_\alpha\,\bar{g}_\beta\,\bar{g}_\gamma \mid S_{k+1} = u]\left(p_{uv}^{(l-k-1)} - \frac{1}{M}\right)\mathrm{E}[\bar{g}_\delta \mid S_l = v]\right| \\ &\leq \frac{1}{M}\sum_{u=1}^{M}\sum_{v=1}^{M}\left|\mathrm{E}[\bar{g}_\alpha\,\bar{g}_\beta\,\bar{g}_\gamma \mid S_{k+1} = u]\right|\left|p_{uv}^{(l-k-1)} - \frac{1}{M}\right|\left|\mathrm{E}[\bar{g}_\delta \mid S_l = v]\right| \\ &\leq K'' e^{-cq}\frac{1}{M}\sum_{u=1}^{M}\sum_{v=1}^{M}\mathrm{E}[|\bar{g}_\alpha\,\bar{g}_\beta\,\bar{g}_\gamma| \mid S_{k+1} = u]\,\mathrm{E}[|\bar{g}_\delta| \mid S_l = v] \\ &\leq K' e^{-cq}\end{aligned}$$

where we used the facts that the state process is ergodic and that the entries of the ν-step transition probability matrix tend to $1/M$ exponentially fast for $q \to \infty$.

Now consider the fourth-order central moments of $\hat{\mathbf{r}}$. We have

$$n\mathrm{E}(\hat{\mathbf{r}}_\alpha - \mathbf{r}_\alpha)(\hat{\mathbf{r}}_\beta - \mathbf{r}_\beta)(\hat{\mathbf{r}}_\gamma - \mathbf{r}_\gamma)(\hat{\mathbf{r}}_\delta - \mathbf{r}_\delta)$$
$$= n^{-3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} \mathrm{E}\, \bar{g}_\alpha(X_i, X_{i+1})\, \bar{g}_\beta(X_j, X_{j+1})\, \bar{g}_\gamma(X_k, X_{k+1})\, \bar{g}_\delta(X_l, X_{l+1}) \tag{4.45}$$

According to (4.42), each of the terms in the sum in (4.45) are bounded. We now divide the terms into groups. First, we divide them into 4! groups depending on the ordering of the indices. We explicitly consider only the group where $i \le j \le k \le l$; a similar result holds for the remaining groups. Now we further divide this group of terms into two smaller groups, one of which has terms that are $O(1)$ and one that has terms that are $O(e^{cn})$. Let q be an integer. Let Σ' denote the sum over all the indices such that $j - i < q$ and $l - k < q$. Let Σ'' denote the remaining terms. The number of terms in Σ' is upper bounded by n^2q^2. We obtain this bound by reasoning as follows. The indices i and k may take at most n values, while the indices j and l may take at most q values. If we take $q = n^{1/4}$, then the number of terms in Σ' grows as $n^{5/2}$, and thus in the sum (4.45) these terms are neglible for large n. The result in (4.44) applies to the terms in Σ'' if we take $q = n^{1/4}$. The number of terms is on the order of n^4; however, because the bound on the size of the individual terms decays exponentially, we see that these terms also become negligible in (4.45) as $n \to \infty$. Hence, we conclude that the fourth-order central moments of $\hat{\mathbf{r}}$ are $o(n)$.

Finally, we know that

$$\theta_n^* = (\mathbf{Z}'\mathbf{C}'\mathbf{W}\mathbf{C}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{C}'\mathbf{W}(\hat{\mathbf{r}}_n - \mathbf{C}\mathbf{Y}\mathbf{p}_y)$$

so that the fourth-order central moments of θ_n^* are a linear combination of the fourth-order central moments of $\hat{\mathbf{r}}$. Therefore, the fourth-order central moments of θ_n^* are also $o(n)$. Q.E.D.

The main results are now stated in the following two theorems.

Theorem 4.7 *Suppose that the underlying state process is ergodic, that* $\mathrm{E}\, g_k(X_1, X_2)^4 \le \Delta < \infty$ *for* $k = 1, 2, \ldots, K$, *and that* $\theta \in \Theta^\circ$. *Let the estimate* $\hat{\theta}_n$ *be obtained as the solution of the minimization problem (4.35). Then the asymptotic covariance matrix for* $\hat{\theta}_n$ *is*

$$\mathbf{V}(\theta) \stackrel{\text{def}}{=} \lim_{n\to\infty} n\mathrm{E}\,(\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)' = (\mathbf{Z}'\mathbf{C}'\mathbf{W}\mathbf{C}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{C}'\mathbf{W}\Sigma\mathbf{W}\mathbf{C}\mathbf{Z}(\mathbf{Z}'\mathbf{C}'\mathbf{W}\mathbf{C}\mathbf{Z})^{-1}$$

where

$$\Sigma = \lim_{n\to\infty} n\mathrm{E}(\hat{\mathbf{r}} - \mathbf{r})(\hat{\mathbf{r}} - \mathbf{r})'$$

is asymptotic covariance matrix of the estimated moment vector $\mathbf{r}$.

Proof. According to Lemma 4.5, $\hat{\theta}_n$ has the same asymptotic covariance matrix as θ_n^*. We obtain θ_n^* by the projection theorem from linear algebra,

$$\begin{aligned}\theta_n^* &= (\mathbf{Z}'\mathbf{C}'\mathbf{W}\mathbf{C}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{C}'\mathbf{W}(\hat{\mathbf{r}} - \mathbf{C}\mathbf{Y}) \\ &= \mathbf{P}(\hat{\mathbf{r}} - \mathbf{C}\mathbf{Y})\end{aligned}$$

where $\mathbf{P} = (\mathbf{Z}'\mathbf{C}'\mathbf{W}\mathbf{C}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{C}'\mathbf{W}$ is the projection matrix onto the linear subspace spanned by the columns of $\mathbf{CZ}$. For the asmptotic covariance matrix $\mathbf{V}(\theta)$, we have

$$\begin{aligned}\mathbf{V}(\theta) &= \lim_{n\to\infty} n\mathrm{E}(\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)' \\ &= \lim_{n\to\infty} n\mathrm{E}(\theta_n^* - \theta)(\hat{\theta}_n^* - \theta)' \\ &= \lim_{n\to\infty} n\mathrm{E}\,\mathbf{P}(\hat{\mathbf{r}} - \mathbf{r})(\hat{\mathbf{r}} - \mathbf{r})'\mathbf{P}' \\ &= \mathbf{P}\Sigma\mathbf{P}'\end{aligned}$$

Q.E.D.

Theorem 4.8 *The asymptotic covariance matrix* $\mathbf{V}(\theta)$ *is bounded by*

$$\mathbf{V}(\theta) \geq (\mathbf{Z}'\mathbf{C}'\Sigma^{-1}\mathbf{C}\mathbf{Z})^{-1} \tag{4.46}$$

and equality holds if $\mathbf{W} = \Sigma^{-1}$. *Inequality in (4.46) is interpreted to mean that the difference between the left- and right-hand sides is positive definite.*

Proof. Let $\mathbf{P} = (\mathbf{Z}'\mathbf{C}'\mathbf{W}\mathbf{C}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{C}'\mathbf{W}$. Then we write

$$\begin{aligned}\mathbf{V}(\theta) - (\mathbf{Z}'\mathbf{C}'\Sigma^{-1}\mathbf{C}\mathbf{Z})^{-1} &= \mathbf{P}\Sigma\mathbf{P}' - (\mathbf{Z}'\mathbf{C}'\Sigma^{-1}\mathbf{C}\mathbf{Z})^{-1} \\ &= [\mathbf{P} - (\mathbf{Z}'\mathbf{C}'\Sigma^{-1}\mathbf{C}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{C}'\Sigma^{-1}]\Sigma[\mathbf{P} - (\mathbf{Z}'\mathbf{C}'\Sigma^{-1}\mathbf{C}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{C}'\Sigma^{-1}]'.\end{aligned}$$

It is obvious that the right-hand side is positive semidefinite. That equality holds for $\mathbf{W} = \Sigma^{-1}$ follows by making the substitution. Q.E.D.

Theorem 4.8 provides us with the best weight matrix in the sense of yielding the smallest asymptotic variance of the estimate. In practice, however, there is a problem applying the result, since one does not know the covariance matrix Σ. Friedlander and Porat suggest that the matrix Σ be replaced by an estimate obtained from the data.

4.5.2 Minimum Discrimination Information Estimation

The minimum discrimination information (MDI) approach to estimation is unique in that it explicitly accommodates a source distribution and a model distribution. The source distribution is the true, unknown distribution of the observations. The model distribution, by contrast, is the assumed parametric distribution, which may be different from the true distribution. This distinction between the source distribution and the model distribution is not explicitly recognized in most other estimation methods.

In short, the MDI method tries to match the model distribution to the source distribution as closely as possible. How close the two distributions are is measured by the discrimination information [9]. Thus, let Q and P be probability distributions with $Q \ll P$ (Q absolutely continuous with respect to P). The discrimination information $D(Q\|P)$ of Q relative to P is defined as

$$D(Q\|P) = \int \log\left[\frac{dQ}{dP}\right] dQ$$

where dP/dQ is the Radon-Nikodym derivative. The discrimination information is still defined even if Q is not absolutely continuous with respect to P. In such case, $D(Q\|P) = \infty$. The discrimination information goes by many names, including cross-entropy, relative entropy, I-divergence, directed divergence, and Kullback-Leibler divergence. If Q and P both have a density function with respect to a measure λ, then the discrimination information is

$$D(Q\|P) = \int \log\left[\frac{q(x)}{p(x)}\right] q(x)\, d\lambda(x).$$

Now suppose that Q^* is the true, unknown source distribution and that P_θ is the model distribution with parameter θ. We cannot discover the true distribution Q^*; however, we can obtain observations $x_1, \ldots, x_n$ and compute the sample moments

$$\hat{r}_k = (n-1)^{-1} \sum_{t=1}^{n-1} g_k(x_t, x_{t+1}) \qquad k = 1, 2, \ldots, K \tag{4.47}$$

as estimates of the true moments

$$r_k = \mathrm{E}_{Q^*}\, g_k(X_1, X_2) \qquad k = 1, 2, \ldots, K \tag{4.48}$$

where the expectation is taken under the distribution Q^*. The functions g_k need not be polynomials; we use the term "moments" in a generic sense.

For large n, the sample moments $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_k)$ should be close to the true moments $\mathbf{r} = (r_1, \ldots, r_k)$; however, in the MDI estimation method, we assume that the two are equal, $\hat{\mathbf{r}} = \mathbf{r}$. This assumption defines the class $\mathcal{Q}$ of distributions that are assumed to contain the true distribution Q^*. Specifically, $\mathcal{Q}$ is defined by

$$\mathcal{Q} = \{Q : \mathrm{E}_Q\, g_k(X_1, X_2) = \hat{r}_k, k = 1, 2, \ldots, k\}.$$

The MDI estimate P_{θ^*} of the model distribution is the distribution that solves the double minimization problem,

$$P_{\theta^*} = \arg\min_{P_\theta} \min_{Q \in \mathcal{Q}} D(Q \| P_\theta).$$

Ephraim, Dembo, and Rabiner [9] describe a method for MDI estimation of HMMs when the output pdfs are multivariate Gaussian. We adapt their approach to our own flavor of HMMs. The vector parameter θ to be estimated comprises the entries of the transition probability matrix, $\theta = \mathbf{p} = (p_{11}, p_{12}, \ldots, p_{MM})$, and the distribution P_θ is the HMM/OS distribution represented by the two-dimensional pdf.

For a given P_θ, the problem of finding a distribution $Q' \in \mathcal{Q}$ to solve $D(Q' \| P_\theta) = \min_{Q \in \mathcal{Q}} D(Q \| P_\theta)$ has been solved by Cziszar in [8]. Cziszar calls this distribution Q' the I-projection of P_θ onto $\mathcal{Q}$ and gives the pdf

$$q'(x, y) = c \exp\left\{\sum_{k=1}^{K} t_k g_k(x, y)\right\} p_\theta(x, y) \tag{4.49}$$

where p_θ is the pdf for P_θ and c and t_k are constants. The constant c is a normalizing constant that may be obtained by integrating the expression. The other constants t_k are obtained by solving the nonlinear system of equations

$$\begin{aligned} \mathbf{r}_k &= \iint g_k(x, y) q'(x, y)\, dx\, dy \\ &= \iint g_k(x, y) \exp\left\{\sum_{l=1}^{K} t_l g_l(x, y)\right\} p_\theta(x, y)\, dx\, dy, \qquad k = 1, 2, \ldots, K \end{aligned} \tag{4.50}$$

The discrimination information is

$$\begin{aligned} D(Q' \| P_\theta) &= \iint q'(x, y) \log\left[\frac{q'(x, y)}{p_\theta(x, y)}\right] dx\, dy \\ &= \log c + \sum_{k=1}^{K} t_k \iint q'(x, y) g_k(x, y)\, dx\, dy \\ &= \log c + \sum_{k=1}^{K} t_k r_k \end{aligned} \tag{4.51}$$

The nonlinear equations (4.50) may be difficult to solve. Ephraim, Dembo, and Rabiner [9] assume that the functions g_k, $k = 1, \ldots, K$ are polynomials, yielding sample moments for $\hat{r}_k$, $k = 1, \ldots, K$. The method they propose requires that a numerical method be used to solve the equations (4.50). We assume that the (x, y) region is partitioned into subregions R_k, $k = 1, 2, \ldots, K$ and take $g_k(x, y) = I_{R_k}(x, y)$, the indicator function for R_k. Then (4.50) becomes

$$\begin{aligned} r_k &= \frac{c}{M} \sum_{i=1}^{M} \sum_{j=1}^{M} p_{ij} e^{t_k} \iint_{R_k} f^{[i]}(x) f^{[j]}(y)\, dx\, dy \\ &= c e^{t_k} F(R_k) \end{aligned}$$

which implies that $t_k = \log r_k - \log F(R_k) - \log c$. Making the substitution into (4.51), we obtain

$$D(Q' \| P_\theta) = \sum_{k=1}^{k} r_k \log r_k - \sum_{k=1}^{k} r_k \log P_\theta(R_k)$$

To minimize $D(Q' \| P_\theta)$ with respect to the parameter θ, we may ignore the first term. Thus we solve

$$\text{maximize} \quad \sum_{k=1}^{K} r_k \log P_\theta(R_k) \tag{4.52}$$

subject to the constraints on the parameters $\theta = \mathbf{p}$. Note that since $P_\theta(R_k)$ is linear function of θ and the log function is concave, the objective function of the maximization problem (4.52) is concave. This property is important when a numerical optimization routine is used, since it guarantees that a unique maximum point exists.

4.5.3 Maximum Likelihood Estimation

Maximum likelihood (ML) estimation is an old stand-by that is a favorite of statisticians because of its asymptotic properties. The method involves solving the maximization problem

$$\begin{aligned} &\text{maximize} && f_\theta(\mathbf{x}) \\ &\text{subject to} && \mathbf{A}\theta = \mathbf{b} \\ & && p_{ij} \geq 0, \quad 1 \leq i, j \leq M \end{aligned} \tag{4.53}$$

where θ is the reduced vector of parameters, $\mathbf{x}$ is the collection of samples, and f_θ is the multivariate pdf when the free parameters are given by θ. The efficient techniques that are available to us for computing the likelihood function and its first derivatives, as presented in

Section 4.3, make ML estimation look attractive to us. With just a little additional trouble, we may also compute the second derivatives, which is a boon when we use a numerical optimization routine.

Baum, *et. al.*, developed an elegant and efficient method for obtaining the maximum likelihood estimate based on a re-estimation procedure. In this method, one starts with an initial guess for the parameters of the model, which we denote by the vector θ_0, and uses this value to re-estimate the parameters based on the samples (observations). The re-estimated parameters θ_1 are then used to re-estimate the parameters θ_2, and so on until $\theta_{t-1} = \theta_t$. The estimated parameters θ may include more than just the transition probabilities. Often the parameters of the output pdfs are included in θ. Some problems with this method have been pointed out in [43].

Unfortunately, we are unable to use the re-estimation procedure because of the constraints that both the rows and the columns of the transition probability matrix sum to one. We can, however, adapt this method to some extent. The result is a possible reduction in the amount of computation needed to solve the maximization problem (4.53). We derive this result now.

The basis for the re-estimation procedure is that the problem of maximizing the likelihood function can be performed by maximizing a sequence of subproblems. Thus the auxiliary function Q is defined as

$$Q(\theta, \theta') = \sum_{\mathbf{s}} f_\theta(\mathbf{x}, \mathbf{s}) \log f_{\theta'}(\mathbf{x}, \mathbf{s})$$

where $\theta = (p_{11}, p_{12}, \ldots, p_{MM})$ and

$$f_\theta(\mathbf{x}, \mathbf{s}) = \frac{1}{M} f^{[s_1]}(x_1) \prod_{t=2}^{n} \left[p_{s_{t-1} s_t} f^{[s_t]}(x_t) \right] \tag{4.54}$$

The sum in (4.54) is taken over all state sequences. The procedure begins by choosing an initial guess θ_0 for the vector θ. At the i stage of the procedure, we obtain θ_i from θ_{i-1} by maximizing $Q(\theta_{i-1}, \theta)$ with respect to θ. The proof that the sequence $\{\theta_i\}$ obtained in this

way monotonically increases the likelihood function is quite simple. We have

$$\begin{aligned} Q(\theta,\theta') - Q(\theta,\theta) &= \sum_{\mathbf{s}} f_\theta(\mathbf{x},\mathbf{s}) \log\left[\frac{f_{\theta'}(\mathbf{x},\mathbf{s})}{f_\theta(\mathbf{x},\mathbf{s})}\right] \\ &\leq \sum_{\mathbf{s}} f_\theta(\mathbf{x},\mathbf{s}) \left[\frac{f_{\theta'}(\mathbf{x},\mathbf{s})}{f_\theta(\mathbf{x},\mathbf{s})} - 1\right] \\ &= \sum_{\mathbf{s}} f_{\theta'}(\mathbf{x},\mathbf{s}) - \sum_{\mathbf{s}} f_\theta(\mathbf{x},\mathbf{s}) \\ &= f_{\theta'}(\mathbf{x}) - f_\theta(\mathbf{x}) \end{aligned}$$

with equality only if $f_{\theta'}(\mathbf{x},\mathbf{s}) = f_\theta(\mathbf{x},\mathbf{s})$ for all sequences $\mathbf{s}$. The maximization subproblem is significantly simpler than the original maximization problem, since we may simplify $Q(\theta,\theta')$ as a function of θ'. Thus,

$$\begin{aligned} Q(\theta,\theta') &= \sum_{\mathbf{s}} f_\theta(\mathbf{x},\mathbf{s}) \log f_{\theta'}(\mathbf{x},\mathbf{s}) \\ &= C(\theta) + \sum_{\mathbf{s}} \sum_{t=2}^{M} f_\theta(\mathbf{x},\mathbf{s}) \log p'_{s_{t-1},s_t} \\ &= C(\theta) + \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{t=2}^{n} \alpha_{t-1}(i) f^{[i]}(x_t)\beta_t(j) p_{ij} \log p'_{ij} \\ &= C(\theta) + \sum_{i=1}^{M}\sum_{j=1}^{M} c_{ij} \log p'_{ij} \end{aligned}$$

With θ fixed, we need not concern ourselves with the first term that does not depend on θ'. Furthermore, after we have computed c_{ij} for $1 \leq i,j \leq M$, we need not recompute these as we iterate in order to find the optimal θ'.

4.6 Simulation Studies

We conclude this chapter with simulation results. The simulations were conceived with two goals in mind. First, we want to see how well the estimation methods perform and to compare the different methods. This goal is pursued in the first set of simulations. Second, we want to see how well the models work for signal discrimination. This issue is studied later.

The three estimation methods we consider are the least squares (LS), minimum discrimination information (MDI), and maximum likelihood (ML). LS and MDI are similar,

in that they both require that we compute a set of statistics $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_{K^2})$ that are used in the objective function to be optimized. These statistics are computed using (4.47), where each g_k is an indicator function, $g_k(x, y) = I_{R_k}(x, y)$. The sets R_k form a partition of the (x,y) region on which the bivariate distribution is defined.

For simplicity, we take the univariate distribution to be uniform on (0,1). This choice for the univariate distribution affects only the ML method, since the LS and MDI methods are invariant under transformations. We form the regions R_k by partitioning the unit square into K^2 subsquares.

The data used in the estimation are generated from a HMM/OS with $M = 5$ and the transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.75 & 0.10 & 0.05 & 0.05 & 0.05 \\ 0.10 & 0.70 & 0.10 & 0.05 & 0.05 \\ 0.05 & 0.10 & 0.70 & 0.10 & 0.05 \\ 0.05 & 0.05 & 0.10 & 0.70 & 0.10 \\ 0.05 & 0.05 & 0.05 & 0.10 & 0.75 \end{bmatrix} \tag{4.55}$$

which yields a positive correlation among the random variables. We use each of the different estimation methods to obtain an estimate $\hat{\mathbf{P}}$ of dimension $M = 5$. Because the parameters of $\mathbf{P}$ produce effects that are averaged out, it is difficult to estimate the individual parameters accurately. To illustrate this, we show the results of the estimation methods from $N = 10000$ samples. From the LS method with $K = 10$, we have

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.762008 & 0.001000 & 0.162032 & 0.001000 & 0.073959 \\ 0.123062 & 0.783537 & 0.001000 & 0.042565 & 0.049836 \\ 0.033571 & 0.094757 & 0.525125 & 0.345547 & 0.001000 \\ 0.034641 & 0.048208 & 0.184214 & 0.597611 & 0.135326 \\ 0.046717 & 0.072499 & 0.127629 & 0.013277 & 0.739879 \end{bmatrix},$$

from the MDI method, again with $K = 10$, we have

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.769735 & 0.001000 & 0.155984 & 0.009981 & 0.063300 \\ 0.072366 & 0.809084 & 0.068733 & 0.001000 & 0.048817 \\ 0.118887 & 0.085052 & 0.366157 & 0.428904 & 0.001000 \\ 0.001458 & 0.001000 & 0.357937 & 0.533775 & 0.105831 \\ 0.037554 & 0.103863 & 0.051189 & 0.026340 & 0.781053 \end{bmatrix},$$

Method	K	N	$D(Q\|\|P)$
LS	5	1000	9.2908e−03
	5	2000	7.3267e−03
	5	5000	5.2714e−03
	5	10000	1.8552e−03
LS	10	1000	9.4841e−03
	10	2000	5.3422e−03
	10	5000	4.1245e−03
	10	10000	9.1457e−04
MDI	5	1000	1.1182e−02
	5	2000	5.9291e−03
	5	5000	3.4571e−03
	5	10000	1.4902e−03
MDI	10	1000	7.0129e−03
	10	2000	4.7435e−03
	10	5000	1.9026e−03
	10	10000	7.7710e−04
ML	n/a	1000	1.9661e−02
	n/a	2000	1.9481e−02
	n/a	5000	1.8251e−02
	n/a	10000	1.8264e−02

Table 4.1: Results of simulations for the esimation methods

and finally, from the ML method, we have

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.798672 & 0.074103 & 0.060979 & 0.065246 & 0.001000 \\ 0.083191 & 0.796156 & 0.033226 & 0.086427 & 0.001000 \\ 0.066876 & 0.027329 & 0.903795 & 0.001000 & 0.001000 \\ 0.050261 & 0.101411 & 0.001000 & 0.846327 & 0.001000 \\ 0.001000 & 0.001000 & 0.001000 & 0.001000 & 0.996000 \end{bmatrix}$$

The smallest entries in any of the estimated transition probability matrices is 10^{-3} because of a constraint placed on the variables in the optimization routine. It is a good idea to bound the entries away from zero, since the number of samples taken for estimation is finite.

We notice that the large diagonal matrix entries are sometimes divided and "shared" with their neighboring entries on the off-diagonal. In order to evaluate the estimation results in a meaningful way, we should average out the parameters in some way. We do this

now by computing the discrimination information $D(Q||P)$ where Q is the true distribution with the transition probability matrix (4.55) and P is the estimated distribution. Note that this is not the same as the objective function for MDI estimation. The results for the estimation under this performance criterion are shown in Table 4.1.

When we look at the estimates $\hat{\mathbf{P}}$ given above, we may be inclined to think that the ML estimate is best, since the diagonal entries are not divided among their neighbors, as in the other estimates. The entries of the ML estimate appear upon a casual glance to be nearest to their true values. The results in Table 4.1, however, indicate that the reverse is true: the ML estimate is the worst. We notice several other results in the table. Taking K larger yields a better estimate for both the LS and the MDI methods. We also notice that the estimation methods behave as they should, with the estimates improving as the number N of samples increases.

Next we look at some discrimination results. Again we take $M = 5$ and the univariate distributions to be uniform. The transition matrices are different under the two hypotheses. Under H_0, we have the transition probability matrix

$$\mathbf{P}_0 = \begin{bmatrix} 0.80 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.80 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.80 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.80 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.80 \end{bmatrix}$$

and under H_1 the matrix

$$\mathbf{P}_1 = \begin{bmatrix} 0.5 & 0.2 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.4 & 0.2 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.4 & 0.2 & 0.1 \\ 0.1 & 0.1 & 0.2 & 0.4 & 0.2 \\ 0.1 & 0.1 & 0.1 & 0.2 & 0.5 \end{bmatrix}$$

These matrice were use to generate data for the estimation methods. Then the estimated matrices were used in discrimination simulations. Figure 4.1 shows that performance of each of these three discriminators for $n = 50$ samples per test. Figure 4.2 shows results for the same discriminators for $n = 100$ samples per test. The plots show the error probabilities P_0 and P_1 of the discriminators parameterized by the LRT threshold. The results indicate that the MDI discriminator performs best, but only slightly better than the LS discriminator.

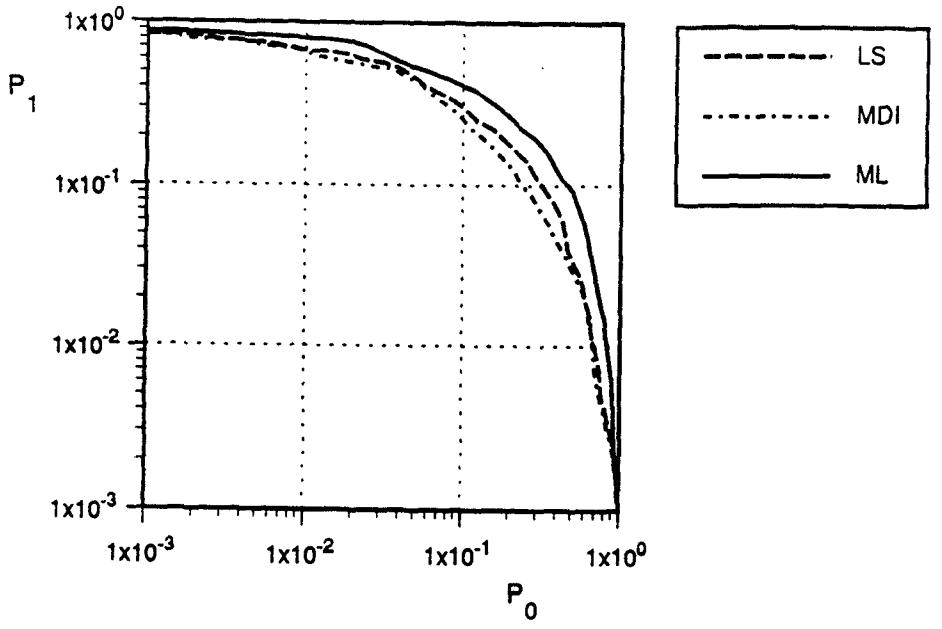


Figure 4.1: Discriminator error probabilities for the LRT discriminator with $N = 50$ samples per test.

shown in the plots corresponds roughly to the ordering shown in Table 4.1. This corroborates the notion that the discrimination information is a good measure of the closeness of two distributions.

4.7 Discussion

In this chapter we have looked at HMMs as a means to obtain multivariate pdfs from arbitrary univariate pdfs. This approach to the use of HMMs is quite different from the usual approach of using HMMs to model nonstationary processes. The multivariate pdfs obtained from HMMs are useful for signal discrimination since they permit us to use an LRT discriminator. Computing the value of the multivariate pdf is fairly simple, using the forward part of the forward-backward algorithm. It is significant that this computation yields the exact value of the pdf and not an approximation as in Chapter 3.

The large number of parameters involved in the HMM multivariate pdf can be a source of difficulty, especially for parameter estimation. On the one hand, the large number of parameters can be a strong point in favor ot eht models, since it permits the flexibility

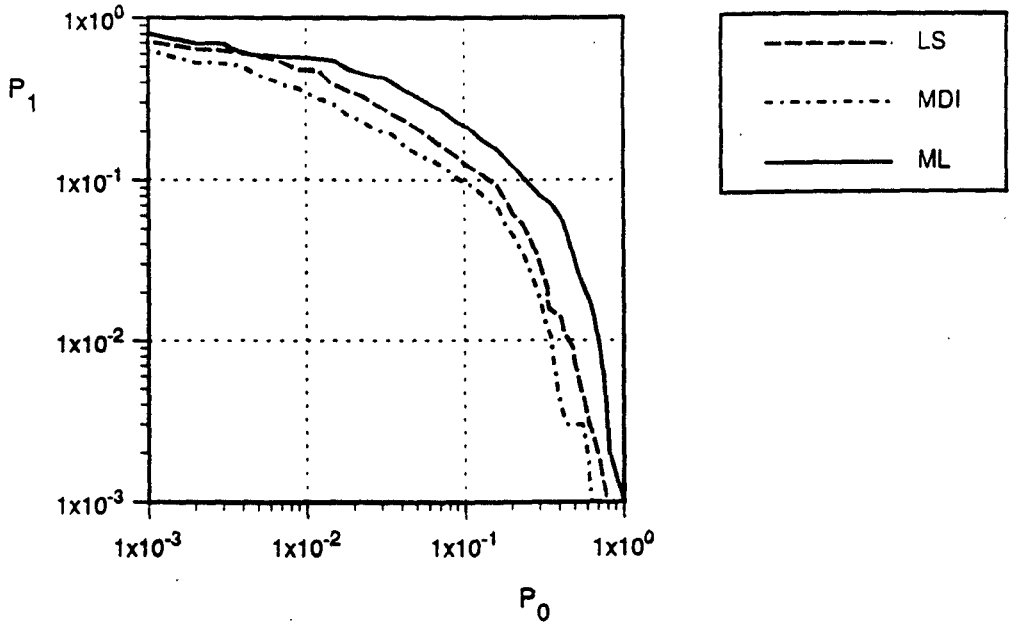


Figure 4.2: Discriminator error probabilities for the LRT discriminator with $N = 100$ samples per test.

of a nonparametric method. On the other hand, trying to fit the model to data by one of the estimation methods discussed here is difficult, requiring that one solve a large scale optimization problem. As an alternative, we can use a parsimonious model with the number of parameters greatly reduced. The parsimonious models have a structure imposed on them, however, that reduces their flexibility.

We looked at three different methods for estimating the parameters of the model. These methods are all similar in that they require that a linearly constrained optimization problem be solved to obtain the estimates. The least squares method is very straightforward, although it can be improved by choosing an appropriate weight matrix. The minimum discrimination method is similar to least squares in that both method require one to compute the values of a collection of statistics that are used in the objective function of the optimization problem. MDI estimation appears to be the best of the three according to our simulation results, and it is the only one of the three that explicitly accommodates a true distribution and a model distribution. Maximum likelihood estimation does not appear to work as well as the other methods. The objective function for the LS and MDI methods is

convex, which guarantees a global optimum point. This is not true for ML—the objective function may have several local optimum points.

Chapter 5

Signal Discrimination Games

5.1 Introduction

Statisticians have occasionally preferred to think of problems in statistical inference as a game involving a decision maker and nature [11]. Nature, as one player in the game, selects a *state of nature*, which is typically a probability distribution or parameter for a distribution. The decision maker, who is the opposing player, observes some data, and then makes a decision about which of several actions to take. Thus, these two players each choose a *strategy*; nature chooses a probability distribution and the decision maker chooses an action. The outcome of the two strategies results in a payoff that is often interpreted as a loss to the decision maker. The problem that statisticians have been concerned with is how the decision maker should choose his action, based on the data he collects, to minimize his expected loss.

The discrimination problem considered in this dissertation may also be viewed as a game involving nature and a decision maker. Nature chooses the hypothesis and the decision maker chooses a discriminator. The error probability of the discriminator may be viewed as the game payoff. In this chapter, however, the interpretation of the discrimination problem as a game is especially appropriate. Here, we consider the discrimination game in which nature is replaced by an adversary of the decision maker. A situation such as this arises in radar target discrimination when the adversary wants to deceive his opponent with a decoy signal. If successful, the decoy signal will be mistaken for a signal from an object that is of interest to the decision maker.

A related topic in signal processing is that of minimax robustness, about which

there is a large volume of literature. The statistical foundation for this area in signal processing was laid by Huber with his two classic papers [21] and [22]. These papers address, respectively, the minimax robust estimation of a location parameter and robust testing of two hypotheses. The work on robust hypothesis testing was later extended by Huber and Strassen in another seminal paper [23]. These three papers laid the foundation for a spate of robustness results in signal processing. A recommended comprehensive survey paper that discusses minimax robustness results in signal processing has been prepared by Kassam and Poor in [27]. Worth mentioning out of the numerous references given there is the paper by Martin and Schwartz [37], which is perhaps the first to address minimax robustness as applied to signal detection. A much more recent paper, and one that perhaps comes closest to addressing the problem we consider here, is one by Jones [25]. This paper,which extends the results of an earlier paper [42] by Morris, addresses the problem of finding a maximin distribution for the noise when the goal is the detection of a stochastic signal in additive noise. The constraints on the noise distribution are similar to some of our constraints, in that the amplitude of the noise is bounded.

The problem we address here differs from minimax robustness in several important ways. One difference concerns the existence of a saddle point. A distinguishing feature of the minimax robustness results is that a saddle point is proved to exist and is explicitly given. In one of the problems considered here, however, no saddle point exists. It seems that a key element necessary for a saddle point to exist is that there be a suitable class of distributions from which the maximin distribution is drawn. Thus, a second important distinguishing feature between minimax robustness and our discrimination game is that our class of distributions is much broader. Finally, whereas the primary focus in robustness is the performance of the discriminator from the point of view of the discriminator designer, our primary focus here is on the performance of the discriminator from the point of view of the signal designer. In the robustness context, one is concerned about how badly the performance is degraded when there is a slight variation from the nominal assumptions about the distribution. In the discrimination game context, however, we are concerned about how best to mimic a particular stochastic signal given various constraints.

In short, the discrimination game pits a discriminator designer against a spurious signal designer in a game. The game is played when the discriminator picks up a signal and makes a decision about whether the signal is the genuine signal of interest or a spurious (decoy) signal. We are concerned about what happens on the average, that is, the error

probability of the discriminator. The discriminator designer wants to minimize the error probability, while the spurious signal designer wants to maximize it. The signal designer must take into consideration certain constraints on the signal. These constraints prohibit him from producing a signal with exactly the same distribution as the genuine signal. As the genuine signal is one that can be measured and recorded by both players, we assume that both players know the distribution of the genuine signal.

Another factor that is important to the spurious signal designer is the average power in the spurious signal. Since the discriminator is required to search for the genuine signal, it is important that the spurious signal have sufficient power to get the attention of the discriminator. A signal that is too weak risks the possibility of being dismissed as pure noise by the discriminator. An increase in the average power of the signal, however, results in a smaller discriminator error probability. Thus an important design consideration is the trade-off between signal power and discriminator error probability. One possibility for dealing with this trade-off, which is the approach considered here, is to establish a lower bound on average power and then obtain the spurious signal with the greatest discriminator error probability subject to the power constraints.

In this chapter, we investigate the signal discrimination game as described above. Constraints on the spurious signal include (1) an upper bound on signal amplitude (which may be interpreted as an upper bound on instantaneous power) and (2) a lower bound on either (a) time-averaged power, or (b) expected power. The case of a lower bound on time-averaged power turns out to be less difficult and less interesting than the case of a lower bound on expected power. In the first case, we explicitly obtain a saddle point pair consisting of a spurious signal distribution and a discriminator decision rule. In the latter case, no saddle point exists and no optimal decision rule is obtained. As the explicit form of the spurious signal distribution depends on the specifics of a particular problem, no explicit spurious signal distribution is given either. Rather, we prove three necessary conditions that are satisfied by the optimal spurious signal distribution. These conditions are sufficient to narrow down the possibilities for the distribution to a collection of distributions parameterized by a single parameter. Thus, one may obtain the optimal spurious signal distribution by solving a one-dimensional maximization problem. This method for obtaining the optimal spurious signal distribution under a lower bound on expected power is illustrated by several examples. Finally, we investigate through simulation studies the design trade-off involving the discriminator error probability and the average power of the

signal.

5.2 Problem Formulation

5.2.1 Players and Strategies

The task of the discriminator is to decide whether the signal it receives is genuine or spurious. For our purposes, it is sufficient to model this task as a binary hypothesis testing problem with H_0 being the hypothesis that the signal is genuine and H_1 the hypothesis that the signal is spurious. The signal received by the discriminator is assumed to be discrete-time, m-dimensional, and strictly stationary. In applications, a discrete-time signal is often obtained by sampling a continuous-time signal. The discriminator collects a sample of size n and makes a decision concerning the authenticity of the signal. Thus we denote the data available to the discriminator by $\mathbf{x} = (x_0, \ldots, x_n)$, where each $x_t = (x_{t1}, \ldots, x_{tm})$ is an m-dimensional vector. Formally stated, the hypotheses are

$$\begin{aligned} H_0 &: \ \mathbf{X} \text{ has a distribution } F_0 \text{ (genuine)} \\ H_1 &: \ \mathbf{X} \text{ has a distribution } F_1 \text{ (spurious)} \end{aligned} \tag{5.1}$$

where $\mathbf{X} = (X_0, \ldots, X_n)$ is a random vector with each component X_t being itself an m-dimensional random vector. While the distribution F_0 of the genuine signal is known to both the discriminator designer and the spurious signal designer, the distribution F_1 is chosen by the spurious signal designer and is unknown to the discriminator designer. We assume that the distribution F_0 is absolutely continuous, having a probability density function f_0. In cases where there is also a density function for F_1, we denote it by f_1.

The discriminator designer, for his strategy, selects a randomized decision rule d such that, when $\mathbf{x}$ is observed, H_1 is decided with probability $d(\mathbf{x})$, and H_0 is decided with probability $1 - d(\mathbf{x})$. Often in discrimination problems d takes only the values 0 or 1, in which case the discriminator is nonrandom, assigning a decision of H_1 or H_0 to each possible observation. The results in this chapter, however, depend heavily on the assumption that d can also take values between 0 and 1. Randomized decision rules are more general than nonrandomized rules and are essential when it is necessary to "interpolate" between two nonrandomized rules d_1 and d_2, for example, by setting $d = (1 - \alpha)d_1 + \alpha d_2$. Essentially, no restrictions are placed on d; any measurable function from $\mathbb{R}^{nm}$ to $[0, 1]$ is permitted. One approach for the discriminator designer is to select a decision rule that is optimal

for a particular distribution F_1. Since the discriminator designer cannot know the true distribution of the spurious signal, a discriminator designed in this manner typically operates in a mismatch situation. This approach has led to useful results in minimax robustness, where it is shown that even though the discriminator operates under mismatched conditions, it is still optimal in a minimax sense. Such a result depends on the existence of a *least favorable* distribution that is universally the worst distribution for all discriminators. For our problem, such a least favorable exists only in the case of a lower bound on time-averaged power.

The likelihood ratio test (LRT) is the one type of decision rule that we are most interested in. To derive this test, we take advantage of a result from real analysis known as the Lebesgue decomposition of a measure [50]. According to this result, for any distribution F_1 we can write $F_1 = F_1^c + F_1^s$, where F_1^c is a sub-probability distribution absolutely continuous with respect to F_0, and F_1^s is a sub-probability distribution singular with respect to F_0. Therefore, there exists a set B such that $F_1^s(B^c) = 0$ and $F_0(B) = 0$; in other words, F_1^c places all of its mass on a set that has F_0-measure 0. On B^c, F_1^c has a density f_1. Thus, we define the likelihood ratio L by

$$L(\mathbf{x}) = \begin{cases} \infty & \mathbf{x} \in B \\ f_1(\mathbf{x})/f_0(\mathbf{x}) & \mathbf{x} \notin B. \end{cases} \tag{5.2}$$

Recall that we assume F_0 to be absolutely continuous with a density function f_0. The LRT decision rule is

$$d(\mathbf{x}) = \begin{cases} 1 & \text{if } L(\mathbf{x}) > \tau \\ \alpha & \text{if } L(\mathbf{x}) = \tau \\ 0 & \text{if } L(\mathbf{x}) < \tau \end{cases} \tag{5.3}$$

where τ is a *threshold* and α is a *randomization parameter*. The family of LRTs, as τ and α vary, is known to be optimal with respect to error probabilities. This property is discussed further below.

While the discriminator designer may choose his strategy, the decision rule d, without restriction, the spurious signal designer's strategy, an nm-dimensional distribution F_1 for the spurious signal, must conform to certain constraints. First, there is an upper bound on the magnitude of the signal, given by

$$F_1\{q_U(X_t) > A_{\max}\} = 0 \tag{5.4}$$

where $q_U(\cdot)$ is a quadratic form. Because of our assumption that the process is stationary,

the bound given by (5.4) is understood to hold for any t. Note that if the signal dimension is $m = 2$, with the components of x_t being in-phase and quadrature components and $q_U(x_t) = x_{t1}^2 + x_{t2}^2$, then (5.4) has the interpretation of a constraint on the instantaneous power or signal amplitude. Hence, we refer to (5.4) generically as an upper bound on the signal amplitude.

A lower bound is also imposed on the signal. We consider three such lower bounds. Analogous to the upper bound (5.4) on the signal amplitude, we may also have a lower bound on the amplitude,

$$F_1\{q_L(X_t) < A_{\min}\} = 0 \tag{5.5}$$

where q_L is also a quadratic form. A more general form of (5.5) is

$$F_1\{q_L(X_0, \ldots, X_N) < D\} = 0 \tag{5.6}$$

where q_L is again a quadratic form. As a particular case, we may have $m = 2$ and

$$q_L(x_0, \ldots, x_N) = \frac{1}{N}\sum_{t=1}^{N}(x_{t1}^2 + x_{t2}^2) \tag{5.7}$$

which has the interpretation of a lower bound on time-averaged power over a window of length N. Hence, we refer to (5.6) generically as a lower bound on time-averaged power. Of course, (5.6) reduces to (5.5) when $N = 1$. Finally, a lower bound may be imposed instead on the expected power of the signal, given by

$$\mathrm{E}_1\, q_L(X_t) \geq E_{\min} \tag{5.8}$$

where E_1 denotes expectation taken under the distribution F_1.

Whereas upper-bound constraints, such as (5.4), typically arise from physical constraints on a transmitter, the lower-bound constraints, such as (5.5), (5.6), and (5.8), are subjectively defined, their purpose being to guarantee a minimum amplitude or power level. A weak signal risks the possibility of being dismissed as pure noise by the discriminator's receiver; hence, a minimum power level is desirable. The constraints on the distribution of the spurious signal define a class $\mathcal{Q}_1$ of distributions from which the spurious signal designer must choose his strategy.

To recapitulate the game formulation stated so far, two players, a discriminator designer and a spurious signal designer, compete against each other in a game. The discriminator designer selects a decision rule d, while the the spurious signal designer selects a

distribution F_1. The class of allowable distributions is defined by constraints that prevent the spurious signal distribution from exactly matching the genuine signal distribution. We assume that the genuine signal distribution is known by both players. We also assume that both players know the class of distributions from which the spurious signal designer must choose his strategy. The game is played when the spurious signal source generates the spurious signal. The discriminator picks up a signal and uses the decision rule to decide whether the signal is genuine or spurious.

5.2.2 The Game Payoff

The final element of the game that must be specified is the payoff function S, which assigns a payoff value to every possible (d, F_1) pair. Loosely speaking, the payoff of the game is the error probability of the discriminator, which the discriminator designer wants to minimize and the spurious signal designer wants to maximize. To be precise, however, we must define the payoff to be a single-valued function of the (conditional) error probabilities P_0 and P_1, which denote the discriminator error probabilities occurring when H_0 and H_1 are true, respectively. For example, P_0 is the conditional probability that the discriminator will decide H_1 given that H_0 is true. To conform to common usage in the statistical literature, we refer to the payoff function as a loss function. The interpretation of S as a loss function is possible only from the discriminator designer's point of view, since the discriminator designer wants to minimize S, while the spurious signal designer wants to maximize S. To avoid confusion, we consistently take the viewpoint of the discriminator designer throughout this chapter. Three loss functions that are commonly used in statistical inference include the Neyman-Pearson, Bayes, and minimax loss functions.

For the Neyman-Pearson loss function, we have $S_N(d, F_1) = P_1$. Notice, though, that with this loss function the problem is trivial unless the class of decision rules is restricted in some way to prevent the use of decision rules such as $\tilde{d} \equiv 1$, since $S(\tilde{d}, F_1) = 0$ for all F_1. Therefore, under the Neyman-Pearson loss function, the class of allowable decision rules is always restricted to a class $\mathcal{D}_\gamma = \{d : P_0(d) \leq \gamma\}$. This rules out the possibility of decision rules like $\tilde{d}$, since $P_0(\tilde{d}) = 1$. Since we assume in our game formulation that F_0 is known, this constraint on the class of allowable decision rules is not inconsistent.

The Bayes loss function, often called *Bayes risk*, for a pair (d, F_1) is defined to be a linear combination of the error probabilities, $S_B(d, F_1) = \pi_0 C_0 P_0 + \pi_1 C_1 P_1$, where

$\pi_i = 1 - \pi_{1-i}$ is the prior probability that H_i is true, and C_i is the cost incurred when an incorrect decision is made and H_i is true. The term "cost," again, is biased toward the discriminator designer's point of view. Bayes risk is, in fact, the expected cost of using the decision rule d when π_i is the probability that H_i occurs. The assumption of prior probabilities π_0 and π_1 is essentially the distinguishing feature of the Bayesian approach in decision theory. What is important for us, though, is just the fact that the Bayes risk is a weighted average of the error probabilities that permits us to establish a total ordering of the decision rules. How the weights are chosen is beyond the scope of this thesis.

The third possibility for a loss function is the so-called minimax loss function $S_M(d, F_1) = \max(C_0 P_0, C_1 P_1)$, where the parameter C_i represents the cost incurred when an incorrect decision is made and H_i is true. Such costs may be subjectively defined. For our purposes, the costs may be interpreted merely as weights reflecting the relative importance of the error probabilities. The minimax loss function may also be interpreted as the Bayes risk when the prior probabilities are unknown, and are therefore assumed to take their worst-case values. In general, when randomized decision rules are permitted, the error probabilities for the minimax decision rule satisfy $C_0 P_0 = C_1 P_1$.

It is well known that the optimal decision rule for the Neyman-Pearson, Bayes, or minimax loss functions is the LRT with τ and α appropriately chosen. (See [46], for example.) This fact is essential to the results that are derived below.

When comparing different spurious signal distributions, it is useful to consider their performance independently of any particular decision rule, if possible. For this purpose, we introduce the concept of a receiver operating characteristic (ROC), a term that stems from the radar literature. Define the partial ordering $\prec$ of decision rules to mean that $d \prec d'$ if $P_0 \leq P_0'$ and $P_1 \leq P_1'$ hold, and if at least one of $P_0 < P_0'$ or $P_1 < P_1'$ holds. Thus $d \prec d'$ means that d is better than d' (from the discriminator designer's point of view). The class $\mathcal{D}$ is defined to be the set of all decision rules such that if $d \in \mathcal{D}$, then there is no d' with $d' \prec d$. It is clear that we need not consider decision rules other than those in $\mathcal{D}$, provided $\mathcal{D}$ is "complete" (see [11]). The set of points $\left[P_0(d), P_1(d)\right]$, as d ranges through $\mathcal{D}$, form a curve called the receiver operating characteristic [57]. Some typical ROCs are shown in Figure 5.1. The reason that LRTs are important to us is that every point (P_0, P_1) of the ROC may be obtained by an LRT with τ and α appropriately chosen. In fact, τ and α parameterize the ROC. Hence, we sometimes refer to points on the ROC as $\left[P_0(\tau, \alpha), P_1(\tau, \alpha)\right]$. In general, whenever randomized decision rules are permitted, the ROC is convex.

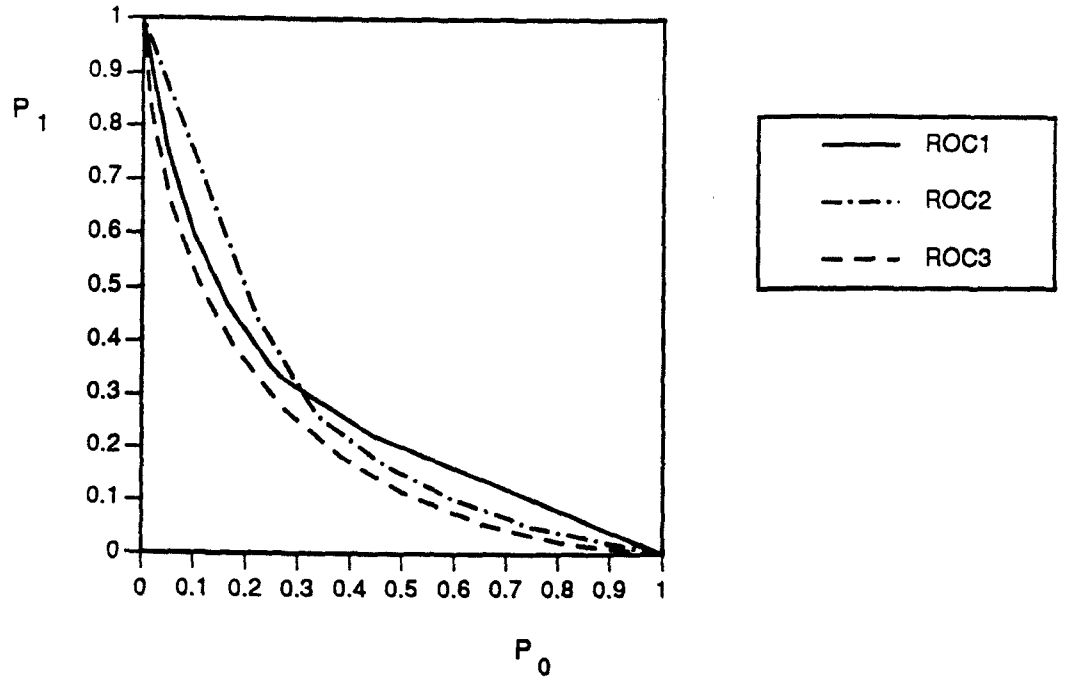


Figure 5.1: Plots of three different receiver operating characteristics (ROCs)

The ROCs provide us with a partial ordering of the distributions F_1. We define the partial ordering $\prec$ to mean that $F_1 \prec F_1'$ whenever the ROC for F_1 lies entirely below the ROC for F_1'. According to this partial ordering, $F_1 \prec F_1'$ implies that F_1 is better than F_1' from the point of view of the discriminator designer. This partial ordering is illustrated in Figure 5.1, where ROC3 is clearly better than either ROC1 or ROC2. Figure 5.1 also demonstrates why the ordering based on ROCs is only a partial ordering. In some places ROC1 is below ROC2, while at other places, it is above. To establish a total ordering of the distributions F_1, we have to specify a particular point on the ROC that is important to us. The various loss functions serve to specify such a point.

In the theory of mathematical games, the goal is usually to obtain strategies that are either maximin or minimax, depending on whether a player's objective is to maximize or minimize the payoff value. A *maximin* solution is defined as a pair (d^M, F_1^M) that satisfies

$$S(d^M, F_1^M) = \max_{F_1} \min_{d} S(d, F_1) \tag{5.9}$$

while a *minimax* solution is defined as a pair (d^m, F_1^m) that satisfies

$$S(d^m, F_1^m) = \min_d \max_{F_1} S(d, F_1) \tag{5.10}$$

A *saddle point* exists when the maximin pair and the minimax pair are the same or, alternatively, when the following inequality holds for all d and F_1:

$$S(\hat{d}, F_1) \leq S(\hat{d}, \hat{F}_1) \leq S(d, \hat{F}_1). \tag{5.11}$$

In (5.11), the pair $(\hat{d}, \hat{F}_1)$ is called a saddle point and $\hat{d} = d^M = d^m$ and $\hat{F}_1 = F_1^M = F_1^m$. It is clearly the goal of the spurious signal designer to employ a maximin density and the goal of the discriminator designer to employ a minimax discriminator.

5.3 Discrimination Game with a Lower Bound on Time-Averaged Power

5.3.1 The Optimal Strategies

We are concerned here with the detection game when the constraints (5.4) and (5.6) are imposed on the spurious signal. The constraint given by (5.5) is included as a special case of (5.6) with $N = 1$. The interpretation we are most interested in is that expressed by (5.7), namely, that of a time-averaged power constraint.

The constraints define a region $B \in \mathbb{R}^{mn}$ which is permitted to have nonzero probability under F_1. Specifically, this region is

$$B = \{\mathbf{x} : \max_{1 \leq t \leq n} q_U(x_t) \leq A_{\max}, \min_{1 \leq t \leq n-N+1} q_L(x_t, \ldots, x_{t+N-1}) \geq D\}. \tag{5.12}$$

The optimal strategies are stated in Theorem 5.1.

Theorem 5.1 *The maximin distribution $\hat{F}_1$ and minimax decision rule $\hat{d}$ form a saddle point of the game. The distribution $\hat{F}_1$ has a density*

$$\hat{f}_1(\mathbf{x}) = \begin{cases} c f_0(\mathbf{x}) & \text{if } \mathbf{x} \in B \\ 0 & \text{otherwise} \end{cases} \tag{5.13}$$

where c is given by

$$c = \left[\int_B f_0(\mathbf{x})\, d\mathbf{x}\right]^{-1}. \tag{5.14}$$

The decision rule $\hat{d}$ is given by

$$\hat{d}(\mathbf{x}) = \begin{cases} \alpha & \text{if } \mathbf{x} \in B \\ 1 & \text{otherwise.} \end{cases} \tag{5.15}$$

Proof. The first inequality in (5.11) holds because the error probabilities of the decision rule do not depend on the shape of the density f_1, but only on the property that $f_1(\mathbf{x}) = 0$, whenever $\mathbf{x} \notin B$. Thus, for any density f_1 which conforms to the lower and upper bounds, there will actually be equality for the first inequality in (5.11). The second inequality holds because $\hat{d}$ is the likelihood ratio test for the densities $\hat{f}_1$ and f_0. Q.E.D.

5.3.2 Case of I.I.D. Observations

Although the result stated in Theorem 5.1 is for correlated observations, there is no significant simplification in the case of i.i.d. observations. If the genuine signal is i.i.d., the spurious signal given by Theorem 5.1 is not i.i.d. in general. However, one may desire to generate the spurious signal as an i.i.d. signal, if the genuine signal is i.i.d. It turns out that the spurious signal is i.i.d. if $N = 1$, so that the constraint (5.6) reduces to (5.5). The density (5.13) becomes a product density

$$\hat{f}_1(\mathbf{x}) = \prod_{t=1}^{n} \hat{f}_1(x_t) \tag{5.16}$$

where

$$\hat{f}_1(x) = \begin{cases} p^{-1} f_0(x) & \text{if } q_U(x) \le A_{\max} \text{ and } q_L(x) \ge A_{\min} \\ 0 & \text{otherwise} \end{cases} \tag{5.17}$$

and p is defined by

$$p = \int_{A_{\min}}^{A_{\max}} f_0(x)\, dx. \tag{5.18}$$

For $N \ge 1$, the decision rule can be defined in terms of a statistic

$$T_n = \sum_{t=1}^{n-N+1} g(x_t, \ldots, x_{t+N-1})$$

where

$$g(x_t, \ldots, x_{t+N-1}) = \begin{cases} 0 & \text{if } (x_t, \ldots, x_{t+N-1}) \in B \\ +\infty & \text{otherwise} \end{cases}$$

and where

$$\hat{d}(\mathbf{x}) = \begin{cases} \alpha & \text{if } T_n = 0 \\ 1 & \text{if } T_n > 0. \end{cases} \tag{5.19}$$

The test statistic T_n resembles the log-likelihood ratio statistic for N-th order Markov processes. When $N = 1$, the test statistic is, in fact, a log-likelihood ratio statistic when

$$g(x_t) = -\log \frac{\hat{f}_1(x_t)}{f_0(x_t)} - \log p \tag{5.20}$$

and the decision rule $\hat{d}$ is defined by (5.19). We must interpret $\log(0) = -\infty$ in (5.20).

5.3.3 Method for Generating the Spurious Signal

In theory, it is quite simple to generate the optimal spurious signal whenever it is possible to generate data from the distribution F_0 of the genuine signal. To generate a vector $\mathbf{x}$ from the multivariate distribution $\hat{F}_1$, first generate a vector $\mathbf{x}$ from the F_0. Then test the vector for acceptance by checking whether $q_L(\mathbf{x}) \geq A_{\min}$ and $q_U(x_t) \leq A_{\max}$, for $t = 1, \ldots, n$. If the conditions hold, then $\mathbf{x}$ is accepted and is the desired vector from $\hat{F}_1$. Otherwise, the vector is rejected. The process is repeated until a vector is finally accepted.

The acceptance-rejection method works well if the probability of acceptance is large enough. In fact, the probability of acceptance is precisely c^{-1}, with c given by (5.14). The number M of vectors that must be generated before one is accepted has a geometric distribution, with the probability of success being c^{-1}. Therefore, the average value of M is c. If c^{-1} is very small, then the average value of M may be too large to make the acceptance-rejection method feasible. If $A_{\min}$ and $A_{\max}$ remain fixed and the dimensionality n is increased, c^{-1} decreases. Thus, if n is large, it is likely that the acceptance-rejection method may be too expensive to use.

If n is too large to make the method feasible, suboptimal methods are possible. Consider, for example, an ARMA model for the genuine signal,

$$X_t + \sum_{i=1}^{L} a_i X_{t-i} = \sum_{j=0}^{M-1} b_j U_{t-j} \tag{5.21}$$

where $\{U_t\}$ is an i.i.d. process and the signal dimensionality is $m = 1$. After each X_t is generated, we check for the upper bound (5.4) and the lower bound (5.6). If the the recently generated X_t is rejected, then we may generate X_t again by generating another U_t and using the same $X_{t-L}, \ldots, X_{t-1}$ and $U_{t-M+1}, \ldots, U_{t-1}$ that were previously used. The method, of course, generates an approximately optimal signal. If i.i.d. data are to be generated, then each component X_t may be generated and tested independently of all other components.

5.4 Discrimination Game with a Lower Bound on Expected Power

Now we consider the discrimination game with the spurious signal constrained by an upper bound (5.4) on amplitude and and a lower bound (5.8) on expected power. The upper bound defines a region R over which the spurious signal distribution F_1 may distribute its mass. Specifically, $R = \{x : q_U(x_t) \leq A_{\max} \text{ for all } t\}$. Recall that $x = (x_1, \ldots, x_n)$, where x_t is an m-dimensional vector for each t. We shall require the assumption that $\sup_{x \in R} q_L(x) = A_{\max}$, an assumption that holds, for example, if $q_L(x) = \prod_{t=1}^{n} q_U(x_t)$.

While the problem involving a lower bound on the time-averaged power has a complete and relatively simple solution, the problem involving a lower bound on expected power is considerably more complicated. Consequently, the results are more difficult to obtain and are not as complete. The results that we do obtain include the specification of necessary conditions that are satisfied by the maximin distribution and a one-parameter characterization of the collection of distributions that satisfy the necessary conditions. To determine the exact maximin distribution therefore requires that one solve a univariate optimization problem. Because of the difficulty of the problem, we do not obtain the minimax decision rule.

Although these results are fairly limited, there are two issues that should be considered. First, although we do not obtain the minimax decision rule for this case, the maximin distribution for the spurious signal seems to be the more valuable result in practice. This is because there are a number of practical considerations that make the specification of the minimax decision rule less useful. For example, in practice a discriminator designer is not safe in assuming that the parameters $A_{\max}$ and $E_{\min}$ of the spurious signal constraints are known. Second, although we do not obtain a general sufficient condition for the maximin distribution, this is a consequence of the nature of the problem and not a consequence of the difficulty of the problem. Since the maximin distribution depends very specifically on the particular distribution F_0, parameters $A_{\max}$ and $E_{\min}$, and loss function, the statement of a general expression for the maximin distribution is not possible. The specification of necessary conditions and a one-parameter family of distributions that satisfy the necessary conditions is the best that can be obtained without specific knowledge about a particular problem.

5.4.1 Sufficient and Necessary Conditions for the Maximin Distribution

In this subsection, we state conditions that must be satisfied by the maximin distribution F_1^M. These necessary conditions limit our search for the maximin distribution to a collection of distributions that are parameterized by a single parameter. Therefore, it is possible to find the exact maximin distribution by solving a univariate maximization problem. We also state and prove a sufficient condition that applies in the special case where Bayes risk is the loss function and the weights on the error probabilities are unequal.

If $E_{\min}$ is too small, then the the maximin distribution is trivial: F_1^M is defined by its probability density function

$$f_1^M(x) = I_R(x)\frac{f_0(x)}{F_0(R)} \tag{5.22}$$

The maximin distribution in this case is nothing more than a version of F_0 truncated at $A_{\max}$. The situation is the same as that of the preceding section with $A_{\min} = 0$. To make the problem interesting, then, we require

$$E_{\min} > \frac{\mathrm{E}_0\, q_L(\mathbf{X})}{F_0(R)} \tag{5.23}$$

This assumption rules out the possibility of (5.22) as the maximin distribution, since it fails to satisfy the lower bound on expected power. When $E_{\min}$ is smaller than the bound in (5.23), then the lower bound on expected power is essentially superfluous. When we assume the bound (5.23) holds, however, probability mass from the distribution in (5.22) must be moved from a region of lower magnitude to a region of higher magnitude in order to satisfy the bound on expected power. The question of how this probability mass should be redistributed is not trivial.

In the following discussions, we frequently compare two distributions $\tilde{F}_1$ and F_1. In our notation, we use a tilde ($\tilde{}$) to indicate quantities that correspond to $\tilde{F}_1$, e.g., $\tilde{\mathrm{E}}_1$ denotes expectation taken under the distribution $\tilde{F}_1$, $\tilde{L}$ denotes the likelihood ratio $\tilde{f}_1/f_0$, and so on. To avoid confusion in our discussions, we consistently take the point of view of the discriminator designer. Therefore, we say that a distribution $\tilde{F}_1$ is *less favorable* than a distribution F_1 when $\tilde{P}_0(\tilde{\tau},\tilde{\alpha}) = P_0(\tau,\alpha)$ implies that $\tilde{P}_1(\tilde{\tau},\tilde{\alpha}) \geq P_1(\tau,\alpha)$. According to this definition, we may have F_1 less favorable than $\tilde{F}_1$, as well as $\tilde{F}_1$ less favorable than F_1, when they have the same ROC. As a stronger condition, we say that $\tilde{F}_1$ is *strictly less favorable* than F_1 when $\tilde{P}_0(\tilde{\tau},\tilde{\alpha}) = P_0(\tau,\alpha) > 0$ and $P_1(\tau,\alpha) > 0$ imply that $\tilde{P}_1(\tilde{\tau},\tilde{\alpha}) > P_1(\tau,\alpha)$. This

condition means that the ROC for F_1 lies strictly below the ROC for $\tilde{F}_1$, except possibly at the endpoints of the ROC.

Ideally, we should like to obtain a distribution F_1 that is universally least favorable in the sense that if $\tilde{F}_1$ is any other distribution, then F_1 is less favorable than $\tilde{F}_1$. After we examine the necessary conditions in detail, it will become clear that such a universally least favorable distribution does not exist. We do not rule out the possibility, however, that a least favorable distribution exists when a specific loss function is assumed, and, in fact, we show later that such a least favorable distribution does exist for the Bayes, minimax, and Neyman-Pearson loss functions. It is important to note, however, that we do not consider the operation of mismatched discriminators, as in the case of minimax robustness. In this sense, our terminology differs from the more common terminology used in the robustness context. Our use of the term least favorable, in this chapter, is synomymous with the term maximin.

The first necessary condition that we derive does not depend on a specific loss function. The condition states that the lower bound on expected power should be satisfied with equality at the maximin distribution. Thus, the condition can be applied to limit our search for the maximin distribution to the class of distributions that satisfy $\mathrm{E}_1\, q_L(X) = E_{\min}$. This makes sense intuitively, since if $\mathrm{E}_1\, q_L(X) > E_{\min}$, then we have some leeway to move around probability mass with the hope of obtaining a less favorable distribution. These intuitive ideas are captured in the following two lemmas. Lemma 5.2 quantifies the idea of "leeway," while Lemma 5.3 shows how to use this leeway to construct a strictly less favorable distribution.

Lemma 5.2 *Let* $\mu = (\mathrm{E}_1\, q_L(X) - E_{\min})/A^n_{\max}$. *Suppose that* $\mathrm{E}_1\, q(X)_L > E_{\min}$ *and that* $\sup_E |F_1(E) - \tilde{F}_1(E)| \leq \mu$. *Then* $\tilde{\mathrm{E}}_1\, q_L(X) \geq E_{\min}$.

Proof. Let B be a set given by the Lebesgue decomposition where $\tilde{F}_1$ is singular with

respect to F_1. Then we have

$$\begin{aligned}
\mathrm{E}_1\, q_L(X) - \tilde{\mathrm{E}}_1\, q_L(X) &= \int_{B^c} q_L(x)\left[1 - \frac{d\tilde{F}_1}{F_1}(x)\right] dF_1(x) + \int_B q_L(x)\, d\tilde{F}_1(x) \\
&= \int_{\{d\tilde{F}_1/dF_1 > 1\}} q_L(x)\left[1 - \frac{d\tilde{F}_1}{F_1}(x)\right] dF_1(x) \\
&\quad + \int_{\{d\tilde{F}_1/dF_1 \le 1\}} q_L(x)\left[1 - \frac{d\tilde{F}_1}{F_1}(x)\right] dF_1(x) \\
&\quad + \int_B q_L(x)\, d\tilde{F}_1(x) \\
&\le A^n_{\max}\left[F_1\{d\tilde{F}_1/dF_1 \le 1\} - \tilde{F}_1\{d\tilde{F}_1/dF_1 \le 1\}\right] + A^n_{\max}\tilde{F}_1(B) \\
&= A^n_{\max}\left[F_1(\{d\tilde{F}_1/dF_1 \le 1\} \cup B) - \tilde{F}_1(\{d\tilde{F}_1/dF_1 \le 1\} \cup B)\right] \\
&\le \mu A^n_{\max}
\end{aligned}$$

Therefore,

$$\tilde{\mathrm{E}}_1\, q_L(X) \ge \mathrm{E}\, q_L(X) - \mu A^n_{\max} = E_{\min}$$

Q.E.D.

Lemma 5.3 *If* $\mathrm{E}_1\, q_L(X) > E_{\min}$, *then there is a distribution* $\tilde{F}_1$ *such that* $\tilde{\mathrm{E}}_1\, q_L(X) \ge E_{\min}$ *and* $\tilde{F}_1$ *is strictly less favorable than* F_1.

Proof. According to Lemma 5.2, we may move around a small amount μ of probability mass to construct $\tilde{F}_1$ without violating the lower bound on expected power.

First, consider the case where F_1 is absolutely continuous. Define $\tilde{F}_1$ on R by its density function

$$\tilde{f}_1(x) = \begin{cases} (1-\mu) f_1(x) & \text{if } L(x) > c \\ c(1-\mu) f_0(x) & \text{if } L(x) \le c \end{cases} \tag{5.24}$$

We need to show that c can be chosen so that $\int \tilde{f}_1(x)\, dx = 1$. Therefore, define $\psi(c) = (1-\mu)^{-1} \int \tilde{f}_1(x)\, dx = F_1\{L > c\} + cF_0(\{L \le c\} \cap R)$. To show that ψ is continuous, we have

$$\begin{aligned}
\psi(c+\epsilon) - \psi(c) &= (c+\epsilon)F_0(\{L \le c+\epsilon\} \cap R) - cF_0(\{L \le c\} \cap R) \\
&\quad - F_1\{L \le c+\epsilon\} + F_1\{L \le c\} \\
&= \int_{\{c < L \le c+\epsilon\} \cap R} \left[c + \epsilon - L(x)\right] f_0(x)\, dx + \epsilon F_0(\{L \le c\} \cap R)
\end{aligned} \tag{5.25}$$

From (5.25), it follows that $0 \le \psi(c+\epsilon) - \psi(c) \le \epsilon$; hence, ψ is continuous. Since $\psi(c+t) - \psi(c) \ge tF_0\{L \le c\}$, it follows that $\psi(c) \to \infty$ as $c \to \infty$. Therefore, since $\psi(0) = 1$, it follows that for any μ between 0 and 1, we may choose c so that $\int_R \tilde{f}_1(x)\,dx = (1-\mu)\psi(c) = 1$. Thus $\tilde{F}_1$ is a valid probability distribution function. By the construction of $\tilde{F}_1$, we have

$$F_1(E) - \tilde{F}_1(E) \le F_1(E) - (1-\mu)F_1(E) = \mu F_1(E) \le \mu$$

Therefore, $\sup_E |F_1(E) - \tilde{F}_1(E)| \le \mu$, and by Lemma 5.2, $\tilde{\mathrm{E}}_1\, q_L(X) \ge E_{\min}$.

Now, for $\tau > c$, we let $\tilde{\tau} = (1-\mu)\tau$ and $\tilde{\alpha} = \alpha$, and then

$$\{L > \tau\} = \{\tilde{L} > \tilde{\tau}\}$$

$$\{L = \tau\} = \{\tilde{L} = \tilde{\tau}\}$$

Therefore,

$$P_0(\tau, \alpha) = F_0\{L > \tau\} + \alpha F_0\{L = \tau\} = F_0\{\tilde{L} > \tilde{\tau}\} + \tilde{\alpha} F_0\{\tilde{L} = \tilde{\tau}\} = \tilde{P}_0(\tilde{\tau}, \tilde{\alpha}) \qquad (5.26)$$

and

$$\begin{aligned} 1 - P_1(\tau, \alpha) &= F_1\{L > \tau\} + \alpha F_1\{L = \tau\} \\ &> (1-\mu)F_1\{L > \tau\} + \alpha(1-\mu)F_1\{L = \tau\} \\ &= \tilde{F}_1\{\tilde{L} > \tilde{\tau}\} + \tilde{\alpha}\tilde{F}_1\{\tilde{L} = \tilde{\tau}\} \\ &= 1 - \tilde{P}_1(\tilde{\tau}, \tilde{\alpha}) \end{aligned}$$

Hence $\tilde{P}_1(\tilde{\tau}, \tilde{\alpha}) > P_1(\tau, \alpha)$.

For $\tau \le c$, we take $\tilde{\tau} = (1-\mu)c$. Now we have

$$\{L > c\} = \{\tilde{L} > \tilde{\tau}\}$$

$$\{L \le c\} = \{\tilde{L} \le \tilde{\tau}\} = \{\tilde{L} = \tilde{\tau}\}$$

The proof that $\tilde{P}_1 > P_1$ for $\tilde{P}_0 = P_0$ is easily proved from the fact that the ROC is convex. For completeness, we work through the proof in detail. For the error probabilities, we have

$$P_0 = F_0\{L > \tau\} + \alpha F_0\{L = \tau\}$$

$$\begin{aligned} \tilde{P}_0 &= F_0\{\tilde{L} > \tilde{\tau}\} + \tilde{\alpha} F_0\{\tilde{L} = \tilde{\tau}\} \\ &= F_0\{L > c\} + \tilde{\alpha} F_0\{L \le c\} \\ &= (1 - \tilde{\alpha})F_0\{L > c\} + \tilde{\alpha} \end{aligned}$$

and

$$1 - P_1 = F_1\{L > \tau\} + \alpha F_1\{L = \tau\}$$

$$\begin{aligned} 1 - \tilde{P}_1 &= \tilde{F}_1\{\tilde{L} > \tilde{\tau}\} + \tilde{\alpha}\tilde{F}_1\{\tilde{L} = \tilde{\tau}\} \\ &= (1 - \tilde{\alpha})\tilde{F}_1\{\tilde{L} > \tilde{\tau}\} + \tilde{\alpha} \\ &= (1 - \tilde{\alpha})(1 - \mu)F_1\{L > c\} + \tilde{\alpha} \end{aligned}$$

Therefore,

$$\begin{aligned} \tilde{P}_1 - P_1 &= F_1\{L > \tau\} + \alpha F_1\{L = \tau\} - (1 - \tilde{\alpha})(1 - \mu)F_1\{L > c\} - \tilde{\alpha} \\ &= [1 - (1 - \tilde{\alpha})(1 - \mu) - \tilde{\alpha}]F_1\{L > c\} + (1 - \tilde{\alpha})F_1\{c \geq L > \tau\} \\ &\quad + (\alpha - \tilde{\alpha})F_1\{L = \tau\} - \tilde{\alpha}F_1\{L < \tau\} \\ &\geq \tau[1 - (1 - \tilde{\alpha})(1 - \mu) - \tilde{\alpha}]F_0\{L > c\} + \tau(1 - \tilde{\alpha})F_0\{c \geq L > \tau\} \\ &\quad + \tau(\alpha - \tilde{\alpha})F_0\{L = \tau\} - \tau\tilde{\alpha}F_0\{L < \tau\} \\ &= \tau[F_0\{L > \tau\} + \alpha F_0\{L = \tau\}] - \tau[(1 - \tilde{\alpha})(1 - \mu)F_0\{L > c\} + \tilde{\alpha}] \\ &> \tau[F_0\{L > \tau\} + \alpha F_0\{L = \tau\}] - \tau[(1 - \tilde{\alpha})F_0\{L > c\} + \tilde{\alpha}] \\ &= \tau(P_0 - \tilde{P}_0) \end{aligned}$$

If $\tilde{\alpha}$ is chosen so that $\tilde{P}_0 = P_0$, then we have $\tilde{P}_1 > P_1$.

Finally, consider the case where F_1 is not absolutely continuous. According to the Lebesgue decomposition of F_1, we may write $F_1 = (1 - \epsilon)F_1^c + \epsilon F_1^s$, where F_1^c is an absolutely continuous probability distribution and F_1^c is a singular probability distribution. Let $\tilde{F}_1 = (1-\epsilon')F_1^c + \epsilon' F_1^s$ with $\epsilon' = \max(0, \epsilon - \mu)$. Let A be the set obtained in the Lebesgue decomposition such that $F_1^s(A) = 1$ and $\lambda(A) = 0$. Then

$$\begin{aligned} \sup_E F_1(E) - \tilde{F}_1(E) &= \sup_E (\epsilon' - \epsilon)F_1^c(E) + (\epsilon - \epsilon')F_1^s(E) \\ &= (\epsilon - \epsilon')F_1^s(A) \\ &= (\epsilon - \epsilon') \\ &\leq \mu \end{aligned}$$

Hence, by Lemma 5.2, $\tilde{\mathrm{E}}_1\, q_L(X) \geq E_{\min}$. Now, we take $\tilde{\tau} = \tau(1 - \epsilon')/(1 - \epsilon)$ and $\tilde{\alpha} = \alpha$,

and then

$$\{L > \tau\} = \{\tilde{L} > \tilde{\tau}\}$$

$$\{L < \tau\} = \{\tilde{L} < \tilde{\tau}\}$$

Thus, (5.26) holds as before. However, we have

$$F_1\{L < \tau\} = (1-\epsilon)F_1^c\{L < \tau\} = (1-\epsilon)F_1^c\{\tilde{L} < \tilde{\tau}\}$$

$$< (1-\epsilon')F_1^c\{\tilde{L} < \tilde{\tau}\} = \tilde{F}_1\{\tilde{L} < \tilde{\tau}\}$$

Similarly, $F_1\{L = \tau\} < \tilde{F}_1\{\tilde{L} = \tilde{\tau}\}$. Therefore, we have

$$\tilde{P}_1 - P_1 = \tilde{F}_1\{\tilde{L} < \tilde{\tau}\} + \tilde{\alpha}\tilde{F}_1\{\tilde{L} = \tilde{\tau}\} + F_1\{L < \tau\} + \alpha F_1\{L = \tau\}$$

$$> 0$$

Hence, $\tilde{P}_1 > P_1$. Q.E.D.

We now consider specific loss functions. Let S denote the generic loss function, which may be Bayes, minimax, or Neyman-Pearson and which remains fixed throughout our discussion. We use S_B when we refer specifically to the Bayes loss function.

For the most part, the different loss functions may be treated uniformly. The exception to this is the Bayes loss function, as we now show. Suppose the weight for one of the error probabilities for the Bayes risk is large relative to the other weight. Then it can happen that the decision rule that achieves the minimum Bayes risk will yield $P_0 = 0$ or $P_1 = 0$. If, for example, one is very strongly inclined to believe that H_0 is true unless there is overwhelming evidence to the contrary, then it is possible that P_0, the probability of deciding H_1 when H_0 is actually true, will be equal to 0 for the optimal decision rule. This situation is illustrated in Figure 5.2. The diagonal lines represent the curves along which the Bayes risk in constant, while the curved line is the ROC for testing F_1 against F_0. As long as the slope of the straight lines is steeper than the slope of the tangent line to the ROC at the point $(0,1)$, the minimum Bayes risk point on the ROC is $(0,1)$. A similar situation occurs when there is a strong bias in favor of H_1. In a real life situation, one can easily imagine how a discriminator designer could bias his decision in favor of H_1, especially if he knows that there are a large number of spurious signal sources around. Also, as one may see clearly in Figure 5.2, this situation can occur more readily when the error probabilities are relatively large. In order to eliminate this special case from interfering with our other results, we state the following sufficient condition.

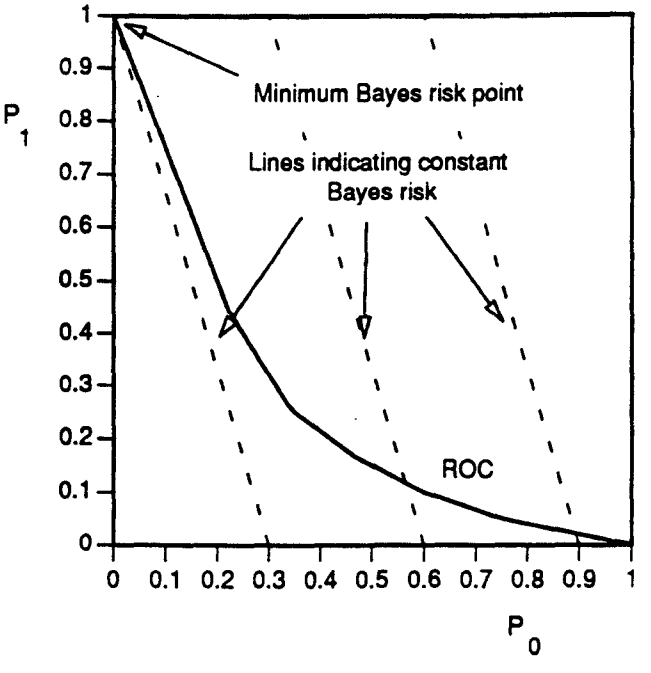


Figure 5.2: Example of a situation where the minimum Bayes risk occurs at a point on the ROC where $P_0 = 0$

Theorem 5.4 *Suppose the distribution F_1 is such that* $\mathrm{E}_1\, q_L(X) \geq E_{\min}$ *and that* $S_B(F_1) = \min(\pi_0 C_0, \pi_1 C_1)$. *Then F_1 is maximin for the Bayes loss function.*

Proof. For any distribution, the risk $\pi_0 C_0$ may be achieved by the trivial decision rule $d \equiv 0$. Similarly, the risk $\pi_1 C_1$ may be achieved by the trivial decision rule $d \equiv 1$. Therefore $\min(\pi_0 C_0, \pi_1 C_1)$ is an upper bound on the Bayes risk. If the minimum Bayes risk for F_1 achieves this upper bound, then we can do no better. Q.E.D.

There are three necessary conditions, as stated in the following three theorems. The first one continues the idea of Lemma 5.3.

Theorem 5.5 *Suppose the distribution F_1 is such that* $\mathrm{E}_1\, q_L(X) \geq E_{\min}$.
(a) If F_1 is maximin for the Bayes loss function and $S_B(F_1) < \min(\pi_0 C_0, \pi_1 C_1)$, *then* $\mathrm{E}_1\, q_L(X) = E_{\min}$.
(b) If F_1 is maximin for the Neyman-Pearson or minimax loss function, then $\mathrm{E}_1\, q_L(X) = E_{\min}$.

Proof. If $\mathrm{E}_1\, q_L(X) > E_{\min}$, then by Lemma 5.3, there exists $\tilde{F}_1$ that is strictly less favorable than F_1, that is, $P_1(\tau, \alpha) < \tilde{P}_1(\tilde{\tau}, \tilde{\alpha})$ whenever $P_0(\tau, \alpha) = \tilde{P}_0(\tilde{\tau}, \tilde{\alpha}) > 0$. Let $\tilde{\tau}'$ and $\tilde{\alpha}'$ yield the values of $\tilde{P}_0$ and $\tilde{P}_1$ where the Bayes risk for $\tilde{F}_1$ is achieved. Let τ' and α' be such that $P_0(\tau', \alpha') = \tilde{P}_0(\tilde{\tau}', \tilde{\alpha}') > 0$. Then $S_B(\tilde{F}_1) = \pi_0 C_0 \tilde{P}_0(\tilde{\tau}', \tilde{\alpha}') + \pi_1 C_1 \tilde{P}_1(\tilde{\tau}', \tilde{\alpha}') > \pi_0 C_0 P_0(\tau', \alpha') + \pi_1 C_0 P_1(\tau', \alpha') \geq S_B(F_1)$

Now let τ' and α' mark the place where $\tilde{F}_1$ achieves its minimax loss, and let τ and α mark the place where F_1 achieves its minimax loss. Since $\tilde{P}_1(\tilde{\tau}, \tilde{\alpha}) > P_1(\tau, \alpha) = P_0(\tau, \alpha) = \tilde{P}_0(\tilde{\tau}, \tilde{\alpha})$, we have $\tilde{P}_1(\tilde{\tau}', \tilde{\alpha}') = \tilde{P}_0(\tilde{\tau}', \tilde{\alpha}') > \tilde{P}_0(\tilde{\tau}, \tilde{\alpha}) = P_0(\tau, \alpha) = P_1(\tau, \alpha)$.

Finally, the result for the Neyman-Pearson loss function follows directly from Lemma 5.3 Q.E.D.

Theorem 5.6 *Suppose the distribution F_1 is such that* $\mathrm{E}_1 q_L(X) \geq E_{\min}$. *Let τ be the threshold for the test that achieves the minimum loss and* $D_1 = \{x : L(x) > \tau\}$.
(a) If F_1 is maximin for the Bayes loss function and $S_B(F_1) < \min(\pi_0 C_0, \pi_1 C_1)$, *then* $q_L(x) = A^n_{\max}$ *for every* $x \in D_1$ *and* $F_0(D_1) = 0$.
(b) If F_1 is maximin for the Neyman-Pearson or minimax loss function, then $q_L(x) = A^n_{\max}$ *for every* $x \in D_1$ *and* $F_0(D_1) = 0$.

Proof. First, note that since F_0 is absolutely continuous, if $q_L(x) = A^n_{\max}$ for every $x \in D_1$ then $F_0(D_1) = 0$. Now, to prove the theorem, we construct a distribution $\tilde{F}_1$ such that $S(\tilde{F}_1) = S(F_1)$. Then we show that if the necessary conditions do not hold, then $\tilde{\mathrm{E}}_1\, q_L(X) > E_{\min}$. According to Theorem 5.5, this means that $\tilde{F}_1$ cannot be maximin, and therefore neither can F_1 be maximin.

Let B be the set of Lebesgue measure 0 where the singular part of F_1 places its mass. Let the absolutely continuous sub-probability measure G be defined by its density function g on B^c, where

$$g(x) = \begin{cases} f_1(x) & \text{if } f_1(x)/f_0(x) < \tau \\ \tau f_0(x) & \text{if } f_1(x)/f_0(x) \geq \tau \end{cases}$$

Let v be a point such that $q_L(v) = A^n_{\max}$. Define $\tilde{F}_1 = G + \mu\delta_v$, where δ_v is a unit point mass at v and $\mu = 1 - G(R)$.

Suppose the loss function is the Bayes loss function S_B. First, we need the following result concerning the Bayes loss function: the minimum Bayes risk is achieved by the LRT with $\tau = \pi_0 C_0/\pi_1 C_1$ and does not depend on the value of α in $[0, 1]$. To prove this statement, let E_1 be the set where we decide H_1. We have

$$\begin{aligned} \pi_0 C_0 P_0 + \pi_1 C_1 P_1 &= \pi_0 C_0 \int_{E_1} f_0(x)\,dx - \pi_1 C_1 \int_{E_1} dF_1(x) + \pi_1 C_1 \\ &= \int_{E_1 - B} [\pi_0 C_0 f_0(x) - \pi_1 C_1 f_1(x)]\,dx - \pi_1 C_1 F_1(E_1 \cap B) + \pi_1 C_1 \end{aligned} \tag{5.27}$$

From (5.27), it follows that to minimize the Bayes risk, we should include x in E_1 if $x \in B$ or if $f_1(x)/f_0(x) > \tau = \pi_0 C_0/\pi_1 C_1$, and exclude x from E_1 if $f_1(x)/f_0(x) < \tau$. If $f_1(x)/f_0(x) = \tau$, then x makes no contribution to the Bayes risk, and therefore α may take any value in $[0, 1]$. Because of this result, we need only consider tests with $\tilde{\tau} = \tau = \pi_0 C_0/\pi_1 C_1$.

The Bayes risk for testing $\tilde{F}_1$ against F_0 is

$$\begin{aligned} S_B(\tilde{F}_1) &= \pi_0 C_0 \tilde{P}_0 + \pi_1 C_1 \tilde{P}_1 \\ &= \pi_0 C_0 \tilde{P}_0 - \pi_1 C_1 (1 - \tilde{P}_1) + \pi_1 C_1 \\ &= \pi_0 C_0 F_0\{\tilde{L} \geq \tau\} - \pi_1 C_1 \tilde{F}_1\{\tilde{L} \geq \tau\} + \pi_1 C_1 \\ &= \pi_0 C_0 F_0\{L \geq \tau\} - \pi_1 C_1 F_1\{L \geq \tau\} + \pi_1 C_1 \\ &= \pi_0 C_0 P_0 + \pi_1 C_1 P_1 \\ &= S_B(F_1) \end{aligned}$$

Now consider the case of the minimax or Neyman-Pearson loss function S. Let $\tilde{\tau} = \tau$ and

$$\tilde{\alpha} = \frac{F_0\{L > \tau\} + \alpha F_0\{L = \tau\}}{F_0\{\tilde{L} = \tau\}} = \frac{F_0\{L > \tau\} + \alpha F_0\{L = \tau\}}{F_0\{L \geq \tau\}}$$

Then

$$\begin{aligned} P_0 - \tilde{P}_0 &= F_0\{L > \tau\} + \alpha F_0\{L = \tau\} - F_0\{\tilde{L} > \tau\} - \tilde{\alpha} F_0\{\tilde{L} = \tau\} \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \tilde{P}_1 - P_1 &= F_1\{L > \tau\} + \alpha F_1\{L = \tau\} - \tilde{F}_1\{\tilde{L} > \tau\} - \tilde{\alpha}\tilde{F}_1\{\tilde{L} = \tau\} \\ &= F_1\{L > \tau\} + \alpha F_1\{L = \tau\} - \Big[F_1\{L > \tau\} - \tau F_0\{L > \tau\}\Big] \\ &\quad - \tau\Big[F_0\{L > \tau\} + \alpha F_0\{L = \tau\}\Big] \\ &= \alpha\Big[F_1\{L = \tau\} - \tau F_0\{L = \tau\}\Big] \\ &= 0 \end{aligned}$$

Therefore $S(\tilde{F}_1) = S(F_1)$.

Finally, we consider $\tilde{\mathrm{E}}_1\, q_L(x)$. Thus,

$$\begin{aligned} \tilde{\mathrm{E}}_1\, q_L(X) - \mathrm{E}_1\, q_L(X) &= -\int_{\{L \geq \tau\}} q_L(x)\, dF_1(x) + \tau \int_{\{L \geq \tau\}} q_L(x) f_0(x)\, dx + \mu A^n_{\max} && \text{(5.28a)} \\ &= \int_{\{L \geq \tau\}} [A^n_{\max} - q_L(x)]\, dF_1(x) && \text{(5.28b)} \\ &\quad - \tau \int_{\{L \geq \tau\}} [A^n_{\max} - q_L(x)] f_0(x)\, dx && \text{(5.28c)} \\ &= \int_{\{L > \tau\}} [A^n_{\max} - q_L(x)]\, dF_1(x) && \text{(5.28d)} \\ &\quad - \tau \int_{\{L > \tau\}} [A^n_{\max} - q_L(x)] f_0(x)\, dx && \text{(5.28e)} \\ &\geq 0 && \text{(5.28f)} \end{aligned}$$

Equality holds in (5.28f) if and only if $F_0\{L > \tau\} = F_0(D_1) = 0$ and $q_L(x) = A^n_{\max}$ for every $x \in D_1$. Therefore, if the necessary conditions are not satisfied, then we have $\tilde{\mathrm{E}}_1\, q_L(x) > E_{\min}$. Q.E.D.

Theorem 5.7 *Suppose the distribution F_1 is such that $\mathrm{E}_1\, q_L(X) \geq E_{\min}$. Let τ be the threshold for the test that achieves the minimum loss and $D_0 = \{x : L(x) < \tau\} \cap R$.*

(a) If F_1 is maximin for the Bayes loss function and $S_B(F_1) < \min(\pi_0 C_0, \pi_1 C_1)$, then there exists a $\gamma \in [0, A^n_{\max}]$ such that D_0 differs from the set $\{x : q_L(x) \le \gamma\}$ by a set of Lebesgue measure 0 and $F_1(D_0) = 0$.

(b) If F_1 is maximin for the Neyman-Pearson or minimax loss function, then there exists a $\gamma \in [0, A^n_{\max}]$ such that D_0 differs from the set $\{x : q_L(x) \le \gamma\}$ by a set of Lebesgue measure 0 and $F_1(D_0) = 0$.

Proof. To prove the theorem, we construct a distribution $\tilde{F}_1$ such that $S(\tilde{F}_1) = S(F_1)$. Then we show that if the necessary conditions do not hold, then $\tilde{\mathrm{E}}_1\, q_L(X) > E_{\min}$. According to Theorem 5.5, this means that $\tilde{F}_1$ cannot be maximin, and therefore neither can F_1 be maximin.

Let $D_* = \{x : L(x) \le \tau\} \cap R$ and define $\tilde{F}_1$ to be the same as F_1 on D_*^c, that is, $\tilde{F}_1(E \cap D_*^c) = F_1(E \cap D_*^c)$ for any set E. Let $u = \operatorname{ess\,sup}_{x \in D_0} q_L(x)$ and $A_\gamma = \{x \in D_* : q_L(x) \le \gamma\}$ where $\gamma \in [0, u]$ is chosen so that $\tau F_0(D_* - A_\gamma) = F_1(D_*)$. To show that such a γ may be found, define the function $\psi(\gamma) = \tau F_0(D_* - A_\gamma) - F_1(D_*)$, which is continuous since F_0 is absolutely continuous. Since $\psi(0) = \tau F_0(D_*) - F_1(D_*) \ge 0$ and $\psi(u) = -F_1(A_\gamma) \le 0$, it follows that $\psi(\gamma) = 0$ for some $\gamma \in [0, u]$. Define $\tilde{F}_1$ on D_* by its density

$$\tilde{f}_1(x) = \begin{cases} 0 & \text{if } x \in A_\gamma \\ \tau f_0(x) & \text{if } x \in D_* - A_\gamma \end{cases}$$

Now, as in the proof of Theorem 5.6, the Bayes risk for testing either $\tilde{F}_1$ or F_1 against F_0 is obtained by the LRT with threshold $\tau = \pi_0 C_0 / \pi_1 C_1$. Thus,

$$\begin{aligned} S_B(\tilde{F}_1) &= \pi_0 C_0 F_0\{\tilde{L} \ge \tau\} + \pi_1 C_1 \tilde{F}_1\{\tilde{L} < \tau\} \\ &= \pi_0 C_0 [F_0\{L > \tau\} + F_0(D_* - A_\gamma)] \\ &= \pi_0 C_0 F_0\{L > \tau\} + \pi_1 C_1 F_1(D_*) \\ &= S_B(F_1) \end{aligned}$$

For the minimax or Neyman-Pearson loss function, let

$$\tilde{\alpha} = \frac{\alpha F_0\{L = \tau\}}{F_0\{\tilde{L} = \tau\}} = \frac{\alpha F_0\{L = \tau\}}{F_0(D_* - A_\gamma)} = \frac{\alpha F_1\{L = \tau\}}{F_1(D_*)} = \frac{\alpha F_1\{L = \tau\}}{F_1\{L \le \tau\}}$$

if $F_0\{L = \tau\} > 0$ and $\tilde{\alpha} = 0$ otherwise. Obviously $0 \leq \tilde{\alpha} \leq 1$. Then,

$$\begin{aligned}\tilde{P}_0(\tau, \tilde{\alpha}) &= F_0\{\tilde{L} > \tau\} + \tilde{\alpha} F_0\{\tilde{L} = \tau\} \\ &= F_0\{L > \tau\} + \alpha F_0\{L = \tau\} \\ &= P_0(\tau, \alpha)\end{aligned}$$

and

$$\begin{aligned}1 - \tilde{P}_1(\tau, \tilde{\alpha}) &= \tilde{F}_1\{\tilde{L} > \tau\} + \tilde{\alpha} \tilde{F}_1\{\tilde{L} = \tau\} \\ &= F_1\{L > \tau\} + \tau^{-1} \tilde{\alpha} F_0\{\tilde{L} = \tau\} \\ &= F_1\{L > \tau\} + \tau^{-1} \alpha F_0\{L = \tau\} \\ &= F_1\{L > \tau\} + \alpha F_1\{L = \tau\} \\ &= 1 - P_1(\tau, \alpha)\end{aligned}$$

Thus, $S(\tilde{F}_1) = S(F_1)$.

Finally, we have

$$\tilde{\mathrm{E}}_1\, q_L(X) - \mathrm{E}_1\, q_L(X) = \int_{D_*} q_L(x)[\tilde{f}_1(x) - f_1(x)]\, dx \tag{5.29a}$$

$$= \int_{D_* - A_\gamma} q_L(x)[\tau f_0(x) - f_1(x)]\, dx - \int_{A_\gamma} q_L(x) f_1(x)\, dx \tag{5.29b}$$

$$\geq \gamma[\tau F_0(D_* - A_\gamma) - F_1(D_* - A_\gamma)] - \gamma F_1(A_\gamma) \tag{5.29c}$$

$$= \gamma[\tau F_0(D_* - A_\gamma) - F_1(D_*)] \tag{5.29d}$$

$$= 0 \tag{5.29e}$$

Equality holds in (5.29c) if and only if $D_0 - A_\gamma$ has Lebesgue measure 0 and $F_1(D_0) = F_1(A_\gamma) = 0$. Therefore, if the necessary conditions are not satisfied, then we have that $\tilde{E}_1\, q_L(X) > E_{\min}$. Q.E.D.

5.4.2 Interpretation of the Necessary Conditions

A couple of examples will serve to illustrate how the necessary conditions may applied to obtain the maximin distribution. As a first example, we consider the univariate probability density function f_0 shown in Figure 5.3. Let $q_L(x) = x^2$. According to the stated necessary conditions, the maximin distribution must have the general form

$$f_1(x) = c I_{(u, A_{\max}]}(x) f_0(x) + h\delta(A_{\max} - x) \tag{5.30}$$

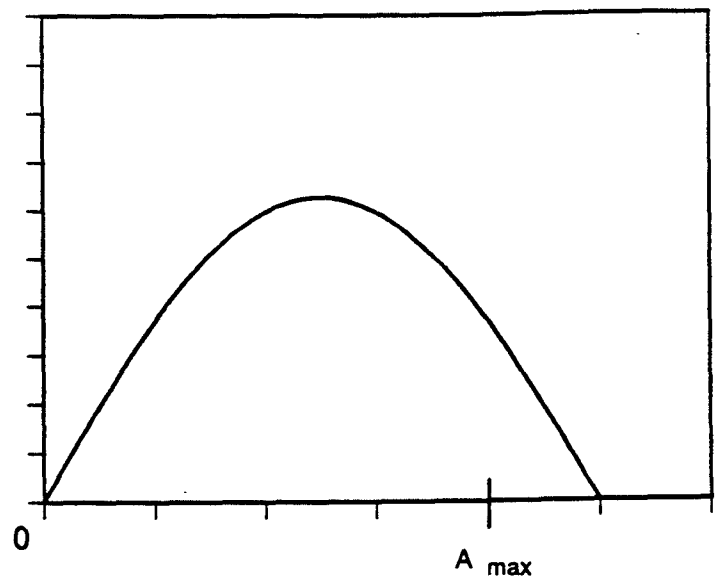


Figure 5.3: Plot of the probability density function f_0.

where δ is a Dirac delta function. A plot of the general form of the density is shown in Figure 5.4. Corresponding to the general form of the maximin density is the general form of the LRT for testing F_1 against F_0,

$$d(x) = \begin{cases} 0 & \text{if } 0 \leq x \leq u \text{ or } x > A_{\max} \\ \alpha & \text{if } u < x < A_{\max} \\ 1 & \text{if } x = A_{\max} \end{cases} \tag{5.31}$$

Note that the decision rule (5.31) involves only one parameter, the randomization parameter α, and that the threshold is unimportant. The entire ROC, a straight line in this case, is parameterized by α. Although there are three parameters appearing in the maximin density, one of them may be eliminated by applying Theorem 5.5, that is, by setting $\mathrm{E}_1\, q_L(X) = E_{\min}$. A second parameter may be eliminated by solving $\int f_1 = 1$. Assume the independent parameter to be u, which takes values in $[0, u_{\max}]$. The plots of the densities for the extreme values of u are shown in Figures 5.5 and 5.6. With u at its minimum value, h takes its maximum value. As u increases to its maximum value, h decreases to 0. This latter case, with $h = 0$, yields the same density as that resulting from a lower bound on amplitude (compare Theorem 5.1).

Given the decision rule (5.31) and the density (5.30), the error probabilities are

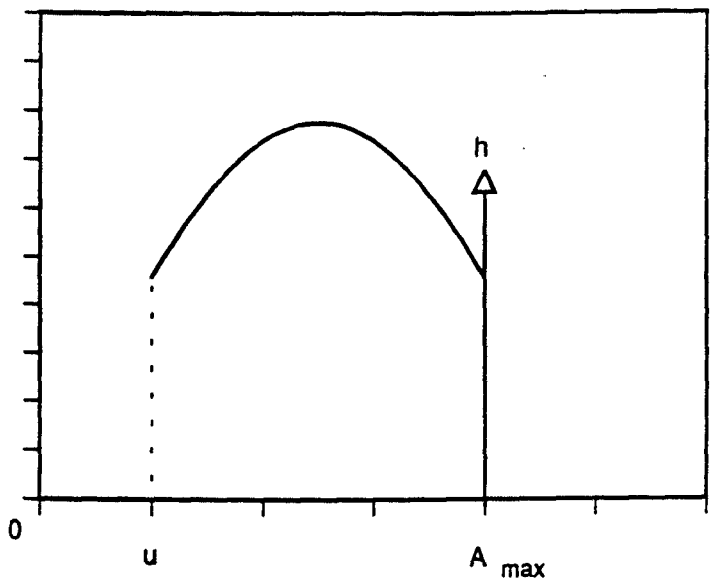


Figure 5.4: Plot of the general form of the maximin density f_1.

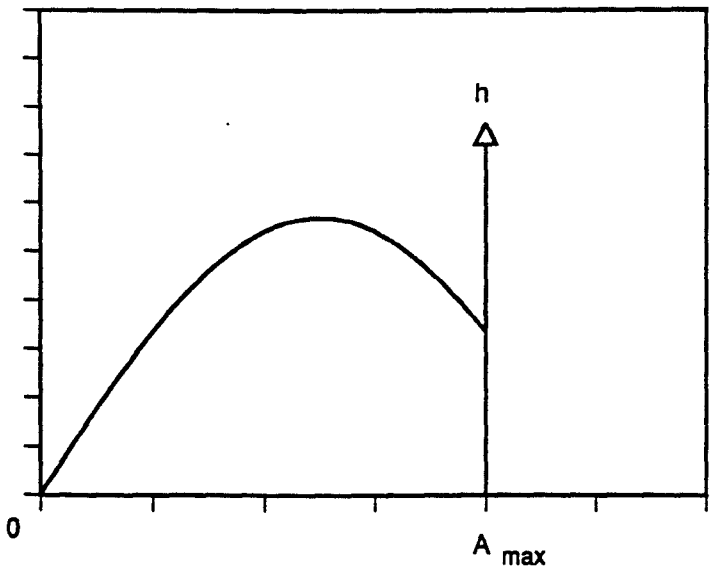


Figure 5.5: Plot of the maximin density f_1 when $u = u_{\min}$.

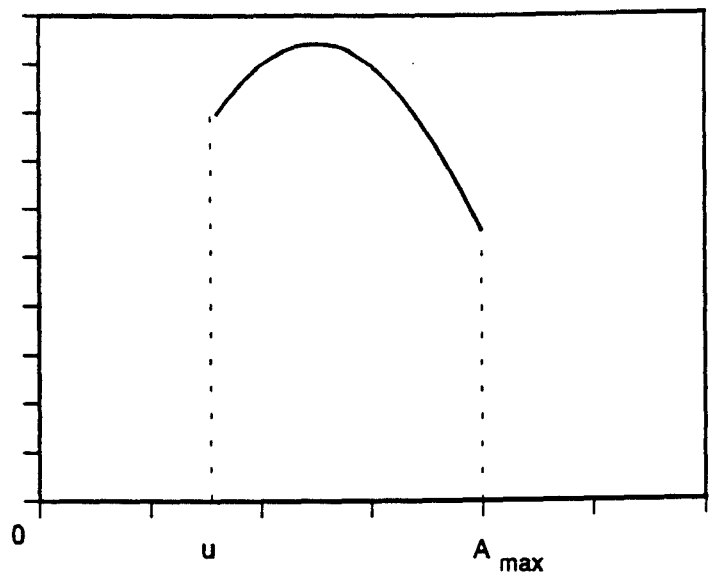


Figure 5.6: Plot of the maximin density f_1 when $u = u_{\max}$.

given by

$$\begin{aligned} P_0 &= \alpha F_0(u, A_{\max}) \\ P_1 &= (1-\alpha) c F_0(u, A_{\max}) \end{aligned} \tag{5.32}$$

These values parameterize the ROC as α varies from 0 to 1. Notice that P_0 can never be 1, since the decision rule always declares H_0 when $x > A_{\max}$, and this happens with positive probability under F_0 and zero probability under F_1. Similarly, P_1 can never be 1 as long as h is positive. The ROCs for the extreme cases of $u = u_{\min}$ and $u = u_{\max}$ are plotted in Figure 5.7. In any case, the decision rule (5.31) is always a completely randomized decision rule, so that the ROC is a straight line. At first glance, it may appear from (5.32) that one should always take $u = 0$, since this would maximize $F_0(u, A_{\max})$. However, upon taking a closer look, we see that c is an implicit function of u, and c increases as u decreases.

As a particular example of how to solve for the maximin distribution F_1, consider the minimax loss function $S_M(d, F_1) = \max(P_0, P_1)$. The value of the loss function is the common value of P_0 and P_1 at the point on the ROC where $P_0 = P_1$. Let $S_M(u)$ denote the minimum value of the loss function for a given F_1 parameterized by the independent parameter u. Setting $P_0 = P_1$ in (5.32) and solving for α, we obtain the common value of

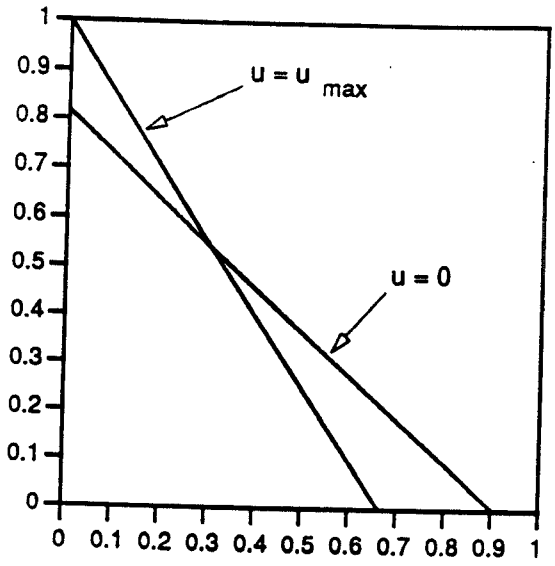


Figure 5.7: ROC for f_1 when $u = u_{\min}$ and $u = u_{\max}$.

P_0 and P_1,

$$S_M(u) = \frac{(A_{\max}^2 - E_{\min})F_0(u, A_{\max})]}{A_{\max}^2 F_0(u, A_{\max}) - G_0(u, A_{\max}) + A_{\max}^2 - E_{\min}} \tag{5.33}$$

where $G_0(u, A_{\max}) = \int_u^{A_{\max}} x^2 f_0(x)\, dx$. Since the function $S_M(u)$ must be maximized with respect to u, we may set its derivative equal to 0 to find any potential maximum point in the interval $(0, u_{\max})$. The derivative is given by

$$\frac{\partial}{\partial u} S_M(u) = \frac{-f_0'(u)(A_{\max}^2 - E_{\min})\left\{A_{\max}^2 - E_{\min} + u^2 F_0(u, A_{\max}) - G_0(u, A_{\max})\right\}}{\left|A_{\max}^2 F_0(u, A_{\max}) - G_0(u, A_{\max}) + A_{\max}^2 - E_{\min}\right|^2}. \tag{5.34}$$

The maximum may also occur at one of the extreme points 0 or $u_{\max}$.

As a second example, we consider a signal of dimension $m = 2$ with f_0 being a bivariate Gaussian density. Let $q_L(\mathbf{x}) = x_1^2 + x_2^2$. We consider the possibility of obtaining a one parameter family of distributions that satisfy the necessary conditions. First, note that the error probabilities are

$$\begin{aligned} P_0 &= \alpha F_0(D_*) \\ P_1 &= (1-\alpha) c F_0(D_*) \end{aligned} \tag{5.35}$$

where D_* is the region where the ratio of f_1 and f_0 is constant. There is also a region D_0 where $f_1 = 0$ and a region D_1 where $F_0(D_1) = 0$ and $F_1(D_1) \geq 0$. Consider now the possible parameters that we may have to deal with. First, there is a parameter h that defines the

amount of probability mass under F_1 that is distributed along the circle $q_L(\mathbf{x}) = A_{\max}^2$. Although it makes no difference in the error probabilities, we might assume for aesthetic reasons, if for no other reason, that this probability mass is distributed uniformly on the circle. Second, there is a parameter u that defines the extent of D_*. We would like u to parameterize D_* in such a way that $F_1[D_*(u)]$ is a decreasing function. We may do this by defining $D_*(u)$ to be the region that maximizes $\int_{D_*(u)} q_L(\mathbf{x})\,d\mathbf{x}$ subject to the constraint that $F_1[D_*(u)] = V(u)$ for some decreasing function $V(u)$. The final parameter in F_1 is the constant c, which plays the role of normalizing the distribution so that it is a probability distribution. In the end, we obtain a loss function S_M like that of (5.33), except that $F_0(D_*)$ replaces $F_0(u, A_{\max})$ and $\int_{D_*(u)} q_L(\mathbf{x}) f_0(\mathbf{x})\,d\mathbf{x}$ replaces $G_0(u, A_{\max})$.

5.5 Simulation Studies

We mentioned briefly in the introduction to this chapter that the purpose of the lower bound constraints on the spurious signal is to increase the average power received by the discriminator. One of the chief concerns of the spurious signal designer is the trade-off between average power received by the discriminator and the probability of error of the discriminator. The spurious signal designer desires to generate the strongest signal possible (up to the physical limit of the transmitter) while keeping the discriminator error probability as high as possible. When the spurious signal source generates continuously at peak transmitter power (a constant signal), an intelligently designed discriminator can discriminate with zero error probability. On the other hand, if no lower bound constraint is placed on the power of the transmitted spurious signal, the discriminator error probability is relatively high. We now present briefly some results which compare the performance of the spurious signal under some distributions that are commonly used in radar modeling, with the average power/error probability trade-off shown explicitly. The error probabilities obtained in the simulation results are those of the optimal matched discriminators.

We assume that the genuine signal has a lognormal distribution. Thus we have $X_i = \exp(\sigma Y_i + \mu)$ where $\{X_i\}$ is the observed genuine signal, and $\{Y_i\}$ is assumed to be a first-order Gaussian autoregressive process with

$$Y_{i+1} = \rho Y_i + \sqrt{1-\rho^2}\epsilon_i \tag{5.36}$$

where $\{\epsilon_i\}$ is an i.i.d. $N(0,1)$ sequence. The parameters are $\mu = 0.8$, $\sigma^2 = 0.25$, and $\rho = 0.9$

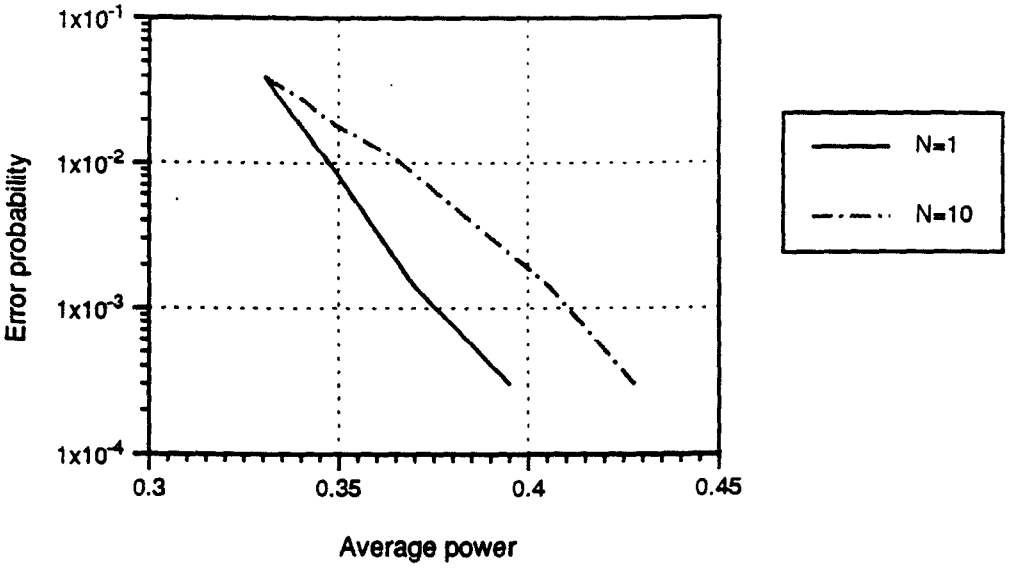


Figure 5.8: Plot of the common value of the error probabilities vs. average power for i.i.d. signals.

for the correlated case and $\rho = 0$ for the i.i.d case. We take $A_{\max} = 4$, which is the 88th percentile of the genuine distribution. Thus, the maximum power which may be transmitted by the spurious signal source is $A_{\max}^2 = 16$. In the included figures, we normalize the average spurious signal power by dividing by the maximum power.

Performance results for the i.i.d. case are shown in Figure 5.8. The discriminator employed in each case is given by Theorem 5.1 with a sample size of $n = 25$. The spurious signal actually generated was the one described in the last part of Section 5.3 for a first order AR process. The ordinates of the plots show the common value of the error probabilities $P_0 = P_1$, obtained by properly choosing the randomization parameter α in the test (5.15). The abscissas show the ratio of the average power actually transmitted to the maximum possible power $A_{\max}^2$. The two plots represent two different lengths for the window over which the power is averaged [see (5.12)]. The case of $N = 1$ involves a lower-bound constraint on the signal amplitude. The curves are parameterized by the value of the lower bound D. The plots show that the longer window length improves the spurious signals performance significantly. Recall that the signal with $N = 1$ is i.i.d., while the other is not.

Figure 5.9 shows a similar situation, but with the genuine signal being correlated.

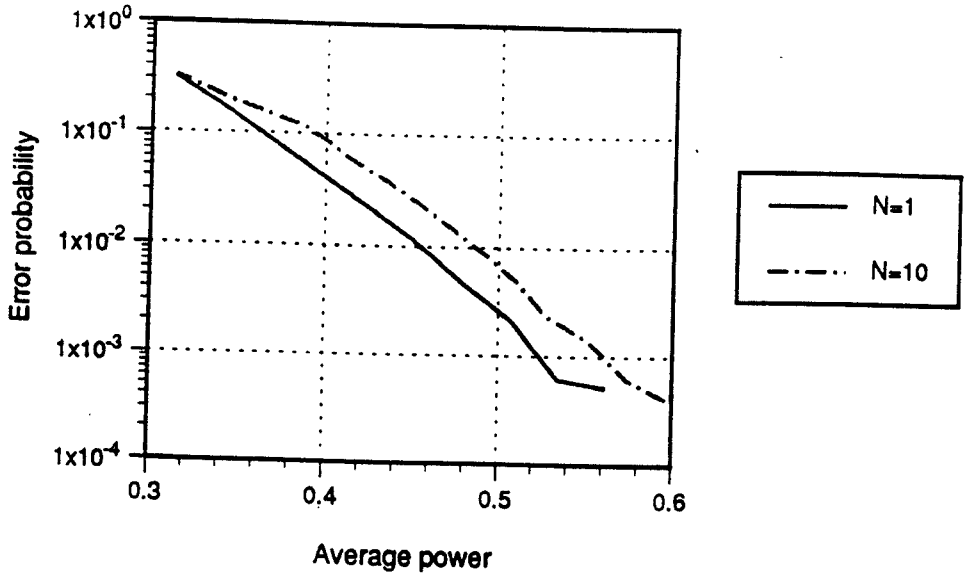


Figure 5.9: Plot of the common value of the error probabilities vs. average power for correlated signals.

The sample size is again $n = 25$. Comparing Figures 5.8 and 5.9 reveals that it is harder for the discriminator to correctly distinguish the signals when the genuine signal is correlated. As before, increasing the window length improves the performance of the spurious signal.

Finally, we look at the performance of the univariate distribution (5.30) obtained under an expected power constraint. For a single sample, the optimal decision rule (LRT) is given by (5.31), and value of the performance measure $\max(P_0, P_1)$ is given by (5.33). For multiple i.i.d. samples, the performance measure under the optimal discriminator is similar to (5.33) and is given by

$$S_M^{(n)}(u) = \frac{(A_{\max}^2 - E_{\min})^n F_0(u, A_{\max}]^n}{(A_{\max}^2 F_0(u, A_{\max}] - G_0(u, A_{\max}])^n + (A_{\max}^2 - E_{\min})^n} \tag{5.37}$$

where n is the sample size. Figure 5.10 shows the value of $S_M^{(n)}$, with a sample size of $n = 25$ and u chosen optimally, plotted as a function of $E_{\min}/A_{\max}^2$; this is shown by the curve labeled EP. The other curves are the same as those shown in Figure 5.8. The univariate density (5.30) obtained under an expected power constraint and used to generate an i.i.d. signal shows a significantly better performance than the densities which were derived under a lower bound on the time-averaged power.

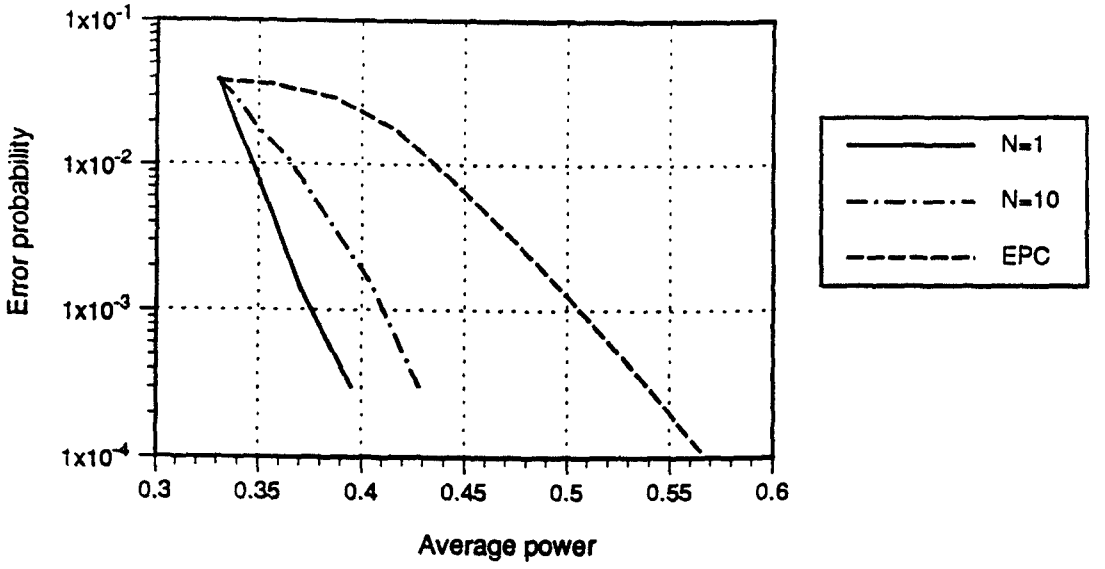


Figure 5.10: Plot of the common value of the error probabilities vs. average power for i.i.d. signals.

5.6 Discussion

This chapter takes a somewhat different perspective into the discrimination problem as compared to the other chapters. While the other chapters focused on methods to improve the performance of a discriminator, the chapter focuses on the problem of how to defeat a discriminator. This problem is somewhat related to the problem considered in minimax robustness. One major difference between our problem and minimax robustness, however, is that the class of allowable distributions from which the unknown distribution may be chosen is much larger. Also, in minimax robustness one is primarily concerned about the performance of the discriminator, while in our problem we are primarily concerned about the performance of the decoy signal.

In summary, we have looked the discrimination problem as a game involving two players, a discriminator designer and a spurious signal designer. While the discriminator designer's goal is to correctly identify a target signal of interest, the spurious signal designer's goal is to deceive the discriminator into accepting a decoy as a true target. Both players know the statistics of the true signal. The spurious signal transmitter, however, cannot exactly reproduce the statistical distribution of the genuine signal because of power

constraints. These constraints are of two different natures. First, there is an upper bound on the peak signal power due to physical constraints on the transmitter. Second, there are subjectively defined constraints that are imposed in order to keep the average power of the signal high, as a signal that is too weak could be dismissed as noise by the discriminator.

We formulated the problem as a mathematical game, with the payoff being Bayes risk, minimax risk, or the error probabilities under a Neyman-Pearson criterion. The results we obtained are divided according to the type of lower bound power constraint that is imposed. When the lower constraint is a bound on time-averaged power, then the minimax discriminator and the maximin signal distribution form a saddle point to the game. We discussed how to generate the spurious signal using an acceptance-rejection method. When the lower constraint is a bound on expected power, then there is not saddle point and the maximin distribution is characterized by necessary conditions. In most cases, these necessary conditions reduce the problem to a one-dimensional problem; the exception occurs only when the payoff is Bayes risk, in which case there is also a sufficient condition. In some cases, depending on the specific problem, we may get the same result for a lower bound on time-averaged power and a lower bound on expected power.

We have looked at some simulation results the showed explicitly the power vs. discriminator error probability for a signal with a lognormal distribution. According to these results, the signal distribution obtained under an expected power constraint showed significantly better performance than the distribution obtained under a time-averaged power constraint for i.i.d. signals. Nevertheless, the signal distribution obtained under a time-averaged power constraint is valuable in practice. It is difficult to generate a signal from the expected power constraint distribution for other than i.i.d. signals. This is not true for the time-averaged power constraint distribution. As we showed, it is not difficult to generate an approximately optimal, correlated signal. Moreover, the signal can be generated indefinitely as a stationary signal, rather than as a signal with fixed length.

Chapter 6

Conclusion

We have looked at several different aspects of the signal discrimination problem. In this last chapter, we summarize our results and discuss the similarities and differences of the techniques that have been presented. Then we discuss some possibilities for future research in this area.

In Chapter 2, we examined discriminators that compute a test statistic as a sum of one-step memory nonlinearities. The optimal nonlinearity was obtained by maximizing a generalized signal-to-noise ratio that depends on the asymptotic means and variances of the test statistic. We also considered the weak-signal detection problem, with the performance measure being the asymptotic relative efficiency. We obtained an incomplete "robustness" result that guarantees a minimal level of performance of the discriminator. Finally, we presented some simulation results that showed a significant improvement of the one-step memory discriminators of the memoryless discriminators.

In Chapter 3, we considered the likelihood ratio test (LRT) discriminator. To use the LRT we derived suitable multivariate pdfs based on a transformation of the χ^2 distribution. By assuming that the χ^2 is generated from underlying Gaussian autoregressive processes, we showed how to simplify the computation of the multivariate χ^2 probability density function (pdf). Our simulation results showed that the transformed χ^2 model can provide an improvement over the transformed Gaussian model for discrimination.

In Chapter 4, we derived a new type of hidden Markov models (HMMs) for stationary processes. These models provide a means to construct a flexible multivariate pdf from an arbitrary univariate pdf. The large number of free parameters in the construction result in a model that tends to be nonparametric in nature. Because of this, parameter

estimation is the key to successful application of the models. It is also possible to reduce the number of parameters by considering parsimonious models. Three different estimation methods, including least squares, minimum discrimination information, and maximum likelihood, were presented and compared. Of these methods, MDI appears to perform best, LS second best, and ML worst.

In Chapter 5, we considered the signal discrimination problem from an entirely different perspective, namely, that of a decoy designer. The idea was to find a signal that would make the discrimination problem the most difficult. The signal was constrained by constraints on its power, including both an upper bound and a lower bound. We showed how to obtain the worst-case signal when there is a lower bound on time-averaged power as well as a lower bound on expected power.

Of the discrimination techniques presented, the one-step memory discriminator of Chapter 2 has the simplest implementation by far. This is its greatest advantage. The main disadvantage of this discriminator is that it is optimal under a generalized SNR and not the error probabilities. Furthermore, this generalized SNR is an asymptotic performance measure that may not be suitable for small or medium sample sizes.

The LRT discriminator based on a transformed χ^2 model is potentially the best discriminator for positive-valued signals, especially for signals that are obtained as the envelope of a complex signal. While the techniques presented in Chapter 3 can make the computation of the multivariate χ^2 pdf manageable in some cases, the expense of computing the pdf remains the limiting factor for these models.

The LRT discriminator based on HMMs are potentially the most flexible. Fitting the models to data is an important problem that has not been solved definitively. This task is perhaps the key to using the models most effectively for discrimination. It is important to note that the models we have considered in Chapter 4 are first-order models. Considering the fact that the underlying state process is Markov, we should probably think in terms of a Markov structure when dealing with the observation process, even though this process is not truly Markov. This first-order structure is definitely a limitation on the HMMs.

Regarding the potential for future research, the HMMs appear the most promising. Fitting the models to data and determining the number of states are both problems for which better solutions can probably be obtained. The investigation of parsimonious models also appears promising. An application area that has not been discussed much in this dissertation, but which is nonetheless important, is Monte Carlo simulation. Both the

transformed χ^2 models and the HMMs are well-suited to this task. The fact that we may compute the exact multivariate pdf for the HMM is a plus, since we may then estimate the probability of certain low probability events via importance sampling.

Appendix A

Rayleigh and Lognormal Probability Density Functions

For the Rayleigh distribution the bivariate joint pdf is

$$f(z,w) = \frac{zw}{\theta^2(1-\rho^2)} \exp\left\{-\frac{z^2+u^2}{2(1-\rho^2)\theta}\right\} I_0\left\{\frac{\rho zw}{(1-\rho^2)\theta}\right\}$$

Furthermore, the third-order joint pdf is

$$f(z,w,v) = \frac{zwv}{\theta^3(1-rho^2)^2} \exp\left\{-\frac{z^2+(1+\rho^2)w^2+v^2}{2(1-\rho^2)\theta}\right\} I_0\left\{\frac{\rho zw}{(1-\rho^2)\theta}\right\} I_0\left\{\frac{\rho wv}{(1-\rho^2)\theta}\right\}$$

Finally, the fourth-order pdf of $(X_1, X_2, X_{l+1}, X_{l+2})$ is

$$f(z,w,v,u) = \frac{zwvu}{\theta^3(1-\rho^2)^2} \exp\left\{-\frac{1}{2(1-\rho^2)\theta}\left[z^2+u^2+\frac{1-\rho^{2l}}{1-\rho^{2(l-1)}}(w^2+v^2)\right]\right\}$$
$$\times I_0\left\{\frac{\rho zw}{(1-\rho^2)\theta}\right\} I_0\left\{\frac{\rho^{l-1}wv}{(1-\rho^{2(l-1)})\theta}\right\} I_0\left\{\frac{\rho vu}{(1-\rho^2)\theta}\right\}$$

Similarly, for the lognormal distribution the bivariate pdf is

$$f(z,w) = \frac{1}{2\pi\theta zw\sqrt{1-\rho^2}} \exp\left\{-\frac{(\log z-\mu)^2+(\log w-\mu)^2-2\rho(\log z-\mu)(\log w-\mu)}{2(1-\rho^2)\theta}\right\}$$

and the third-order pdf is

$$f(z,w,v) = \frac{1}{zwv(2\pi\theta)^{3/2}(1-\rho^2)} \exp\Big\{-\frac{1}{2(1-\rho^2)\theta}\Big[(\log z-\mu)^2+(\log v-\mu)^2\Big]$$
$$+(1+\rho^2)(\log w-\mu)^2-2\rho\Big[(\log z-\mu)(\log w-\mu)+(\log w-\mu)(\log v-\mu)\Big]\Big\}$$

Finally, the fourth-order joint pdf of $(X_1, X_2, X_{l+1}, X_{l+2})$ is given by

$$f(z,w,v,u) = \frac{1}{(2\pi\theta)^2 zwvu(1-\rho^2)\sqrt{1-\rho^{2(l-1)}}} \exp\Bigg\{-\frac{1}{2(1-\rho^2)\theta}\Bigg[(\log z-\mu)^2+(\log u-\mu)^2$$
$$+\frac{1-\rho^{2l}}{1-\rho^{2(l-1)}}\Big[(\log w-\mu)^2+(\log v-\mu)^2\Big]$$
$$-2\rho\Big[(\log z-\mu)(\log w-\mu)+(\log v-\mu)(\log u-\mu)\Big]$$
$$2\frac{\rho^{l-1}(1-\rho^2)}{1-\rho^{2(l-1)}}(\log w-\mu)(\log v-\mu)\Bigg]\Bigg\}$$

Appendix B

Cauchy Probability Density Functions

We start with the Gaussian bivariate, third-order, and fourth-order pdfs, namely

$$f_G(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[x^2 - 2\rho xy + y^2\right]\right\}$$

$$f_G(u,x,y) = \frac{1}{(2\pi)^{3/2}\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[u^2 + (1+\rho^2)x^2 + y^2 - 2\rho(ux+xy)\right]\right\}$$

$$f_G(u,v,x,y) = \frac{1}{(2\pi)^2(1-\rho^2)\sqrt{1-\rho^{2k-2}}} \exp\left\{\frac{\mathbf{z}'\Lambda\mathbf{z}}{2(1-\rho^2)}\right\}$$

where

$$\Lambda = \begin{bmatrix} 1 & -\rho & 0 & 0 \\ -\rho & s & -\rho & 0 \\ 0 & -r & s & -\rho \\ 0 & 0 & -\rho & 1 \end{bmatrix}$$

and $s = (1-\rho^{2k})/(1-\rho^{2k-2})$ and $r = \rho^{k-1}(1-\rho^2)/(1-\rho^{2k-2})$.

Suppose X is Gaussian and that $\Phi(x) = (2\pi)^{-1/2}\int_{-\infty}^{x} e^{-t^2/2}\,dt$. Then the transformation $X = \psi(Y)$ where

$$\psi(x) = \Phi^{-1}\left[\frac{1}{\pi}\tan^{-1}x + \frac{1}{2}\right]$$

implies that Y is Cauchy distributed. The inverse of ψ is

$$\psi^{-1}(x) = \tan\left(\left[\Phi(x) - \frac{1}{2}\right]\pi\right)$$

The first and second derivatives of ψ are

$$\psi'(x) = \sqrt{\frac{2}{\pi}} \frac{\exp\left[\psi(x)^2/2\right]}{1+x^2}$$

and

$$\psi''(x) = \psi(x)|psi'(x)|^2 - \frac{2x}{1+x^2}\psi'(x)$$

Furthermore the bivariate, third-order, and fourth-order pdfs are given by

$$f(x,y) = f_G\left[\psi(x),\psi(y)\right]\psi'(x)\psi'(y)$$

$$f(u,x,y) = f_G\left[\psi(u),\psi(x),\psi(y)\right]\psi'(u)\psi'(x)\psi'(y)$$

and

$$f(u,v,x,y) = f_G\left[\psi(u),\psi(v),\psi(x),\psi(y)\right]\psi'(u)\psi'(v)\psi'(x)\psi'(y)$$

Appendix C

Quadrature Formulas for Integrating on the Surface of the Unit Hypersphere

The unit hypersphere in d dimensions is the set of points $S = \{\mathbf{x} : x_1^2 + \cdots + x_d^2 = 1\}$. We are interested in quadrature formulas that can approximate the integral

$$\int_S f(\mathbf{x})\, d\sigma$$

where $d\sigma$ is an element of area on the surface S. By a quadrature formula, we mean a set $\{(\mathbf{x}_i, w_i),\ 1 \le i \le N\}$ of node and weight pairs such that

$$\int_S f(\mathbf{x})\, d\sigma \approx \sum_{i=1}^{N} f(\mathbf{x}_i) w_i. \tag{C.1}$$

A useful measure of the accuracy of a particular quadrature formula is given by its degree. By definition, a quadrature formula is of degree m if the approximation (C.1) is exact whenever f is a polynomial of degree m or less.

The quadrature formulas given here are symmetric. This means that if $\mathbf{x} = (x_1, \ldots, x_d)$ is a node, then so is $(\pm x_{\sigma_1}, \ldots, \pm x_{\sigma_d})$, where $(\sigma_1, \ldots, \sigma_d)$ is any permutation of the integers $1, 2, \ldots, d$.

The formulas for $d = 2$ are simple and may be obtained from closed form expressions. Because of symmetry, the number N of nodes must be even. Then the ith node is $\mathbf{x}_i = (\cos 2\pi i/N, \sin 2\pi i/N)$ and the corresponding weight is $w_i = 2\pi/N$.

The procedure for obtaining the nodes and weights for a quadrature formula for $d > 2$ is described in [28]. In short, this involves solving a system of nonlinear equations obtained by setting the right and left hand sides of (C.1) equal when f is polynomial chosen out of a basis set for the mth degree polynomials. Patrick Keast computed the weights and node presented below for $d = 3$.

Because of symmetry, it is not necessary to list all of the nodes. The nodes may be divided into equivalence classes and it is necessary only to list one node, called a node generator, out of each class. For example, if $\mathbf{x} = (x_1, \ldots, x_d)$ is given as a node generator, then the other nodes in the class nodes of the form $(\pm x_{\sigma_1}, \ldots, \pm x_{\sigma_d})$. All of the nodes in a class have the same weight.

For each quadrature formula, we give the number of nodes, the degree, the node generators, and the corresponding weights.

Degree	# Nodes	Generators			Weights
3	6	1.0000000000	0.0000000000	0.0000000000	2.0943951024
5	14	1.0000000000	0.0000000000	0.0000000000	0.8377580410
		0.5773502692	0.5773502692	0.5773502692	0.9424777961
7	26	1.0000000000	0.0000000000	0.0000000000	0.5983986007
		0.7071067812	0.7071067812	0.0000000000	0.4787188805
		0.5773502692	0.5773502692	0.5773502692	0.4039190555
9	38	1.0000000000	0.0000000000	0.0000000000	0.1196797201
		0.4597008434	0.8880738340	0.0000000000	0.3590391604
		0.5773502692	0.5773502692	0.5773502692	0.4039190555
11	50	1.0000000000	0.0000000000	0.0000000000	0.1595729602
		0.7071067812	0.7071067812	0.0000000000	0.2836852625
		0.5773502692	0.5773502692	0.5773502692	0.2650718801
		0.3015113446	0.3015113446	0.9045340337	0.2535056109
13	78	1.0000000000	0.0000000000	0.0000000000	0.5571838151e−01
		0.3337005380	0.9426791347	0.0000000000	0.1886150063
		0.4394838395	0.4394838395	0.7833951172	0.1956786568
		0.7011707417	0.7011707417	0.1293026753	0.1253755170
15	86	1.0000000000	0.0000000000	0.0000000000	0.1450663274
		0.3742430391	0.9273306572	0.0000000000	0.1484377867
		0.5773502692	0.5773502692	0.5773502692	0.1500915882
		0.6943540066	0.6943540066	0.1890635529	0.1492445169
		0.3696028465	0.3696028465	0.8525183117	0.1396193608
17	110	1.0000000000	0.0000000000	0.0000000000	0.4810746585e−01
		0.8781589106	0.4783690288	0.0000000000	0.1218309174
		0.5773502692	0.5773502692	0.5773502692	0.1230717353
		0.1851156353	0.1851156353	0.9651240351	0.1031917341
		0.3956894731	0.3956894731	0.8287699813	0.1205802490
		0.6904210484	0.6904210484	0.2159572918	0.1249450969
19	146	1.0000000000	0.0000000000	0.0000000000	0.7535190013e−02
		0.7071067812	0.7071067812	0.0000000000	0.9265184700e−01
		0.5773502692	0.5773502692	0.5773502692	0.9061000834e−01
		0.1574676672	0.1574676672	0.9748886437	0.9518264418e−01
		0.4174961228	0.4174961228	0.8070898184	0.8487112439e−01
		0.6764410400	0.6764410400	0.2912988822	0.8942676055e−01
		0.1403553812	0.4493328323	0.8822700113	0.8785259468e−01

Table C.1: Quadrature formulas for the surface of a sphere.

Bibliography

[1] Lalit R. Bahl, Peter F. Brown, Peter B. de Souza, and Robert L. Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 49–52, 1986.

[2] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41(1):164–171, 1970.

[3] Martin L. Brady, Raghu Raghavan, and Joseph Slawny. Back propagation fails to separate where perceptrons succeed. *IEEE Trans. Circuits and Systems*, 36(5):665–674, 1989.

[4] Stephen P. Bruzzone and M. Kaveh. Information tradeoffs in using the sample autocorrelation function in Arma parameter estimation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP–32(4):701–715, 1984.

[5] James A. Bucklew. *Large Deviations Techniques in Decision, Simulation, and Estimation.* Wiley, New York, 1990.

[6] Ben Cantrell. Detection of signals in non-gaussian correlated noise derived from cauchy processes. Technical Report 9086, U.S. Naval Research Lab, Washington, D.C., 1987.

[7] Jack Capon. On the asymptotic efficiency of locally optimum detectors. *IRE Trans. Inform. Theory*, IT-7:67–71, 1961.

[8] I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3(1):146–158, 1975.

[9] Yariv Ephraim, Amir Dembo, and Lawrence R. Rabiner. A minimum discrimination information approach for hidden Markov modeling. *IEEE Trans. Inform. Theory*, IT-35(5):1001–1013, 1989.

[10] Yariv Ephraim, Hanoch Lev-Ari, and Robert M. Gray. Asymptotic minimum discrimination information measure for asymptotically weakly stationary processes. *IEEE Trans. Inform. Theory*, IT-34(5):1033–1040, 1988.

[11] Thomas S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press, New York, 1967.

[12] Benjamin Friedlander and Boaz Porat. Performance analysis of parameter estimation algorithms based on high-order moments. *Int. J. Adaptive Contr. Signal Processing*, 3:191–229, 1989.

[13] Benjamin Friedlander and Boaz Porat. Asymptotically optimal estimation of MA and ARMA parameters of non-Gaussian processes from high-order moments. *IEEE Trans. Automat. Control*, AC-35(1):27–35, 1990.

[14] William A. Gardner. *Statistical Spectral Analysis—A Nonprobabilistic Theory.* Prentice Hall, Edgewood Cliffs, N.J., 1988.

[15] Georgios B. Giannakis and Michail K. Tsatsanis. Signal detection and classification using matched filtering and higher-order statistics. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-38(7):1284–1296, 1990.

[16] Philip E. Gill, Walter Murray, and Margaret H. Wright. *Practical Optimization.* Academic Press, New York, 1981.

[17] Paul R. Gorman and Terrence J. Sejnowski. Learned classification of sonar targets using a massively parallel network. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-36(7):1135–1140, 1988.

[18] Robert M. Gray, Jr. Augustine H. Gray, Guillermo Rebolledo, and John E. Shore. Rate-distortion speech coding with a minimum discrimination information distortion measure. *IEEE Trans. Inform. Theory*, IT-27(6):708–721, 1981.

[19] Don R. Halverson and Gary L. Wise. Discrete-time detection in ϕ-mixing noise. *IEEE Trans. Inform. Theory*, IT-26:189–198, 1980.

[20] Melvin J. Hinich and Gary R. Wilson. Detection of non-gaussian signals in non-gaussian noise using the bispectrum. *IEEE. Trans. Acoust.. Speech, Signal Processing*, ASSP–38(7):1126–1131, 1990.

[21] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Stat.*, 35:73–101, 1964.

[22] Peter J. Huber. A robust version of the probability ratio test. *Ann. Math. Stat.*, 36:1753–1758, 1965.

[23] Peter J. Huber and Volker Strassen. Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Statistics*, 1:251–263, 1973.

[24] Mark E. Johnson. *Multivariate Statistical Simulation.* Wiley, New York, 1987.

[25] Lee K. Jones. A stochastic multisample extension of Morris's robust detector in bounded amplitude noise. *IEEE. Trans. Inform. Theory*, IT–34(5):973–978, 1988.

[26] B. H. Juang, Stephen E. Levinson, and M. M. Sondhi. Maximum likelihood estmination for multivariate mixture observations of Markov chains. *IEEE. Trans. Inform. Theory*, IT–32(2):307–309, 1986.

[27] Saleem A. Kassam and H. Vincient Poor. Robust techniques for signal processing: A survey. *Proc. IEEE*, 73(3):433–481, 1985.

[28] Patrick Keast. Cubature formulas for the surface of the sphere. *J. Comp. Appl. Math.*, 17:151–172, 1987.

[29] George Kimeldorf and Allan Sampson. One-parameter families of bivariate distribution with fixed marginals. *Comm. in Stat.*, 4(3):293–301, 1975.

[30] George Kimeldorf and Allan Sampson. Uniform representations of bivariate distributions. *Comm. in Stat.*, 4(7):617–627, 1975.

[31] Anthony LaVigna. *Nonparametric Classification using Learning Vector Quantization.* PhD thesis, University of Maryland, 1989.

[32] A. J. Lawrance and P. A. W. Lewis. Modelling and residual analysis of nonlinear autoregressive time series in exponential variables. *J. Roy. Statist. Soc., Ser. B*, 47(2):165–202, 1985.

[33] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell System Technical Journal*, 62(4):1035–1074, 1983.

[34] Shing Ted Li and Joseph L. Hammond. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Trans. on Syst. Man Cybern.*, SMC–5(5):557–561, 1975.

[35] Louis A. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Trans. Inform. Theory*, IT–28(5):729–734, 1982.

[36] John Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63(4):561–580, 1975.

[37] Douglas R. Martin and Stuart C. Schwartz. Robust detection of a known signal in nearly Gaussian noise. *IEEE Trans. Inform. Theory*, IT–17(1):50–56, 1971.

[38] Andrew B. Martinez, Peter F. Swaszek, and John B. Thomas. Locally optimal detection in multivariate non-Gaussian noise. *IEEE Trans. Inform. Theory*, IT-30(6):815–822, 1984.

[39] Neri Merhav. Universal classification for hidden Markov models. *IEEE Trans. Inform. Theory*, IT-37(6):1586–1594, 1991.

[40] James H. Miller and John B. Thomas. Detectors for discrete-time signals in non-Gaussian noise. *IEEE Trans. Inform. Theory*, IT–18(2):241–250, 1972.

[41] Kenneth S. Miller. *Multidimensional Gaussian Distributions.* Wiley, New York, 1964.

[42] Joel M. Morris. On single-sample robust detection of known signals with additive unknown-mean amplitude-bounded random interference—Part II: The randomized decision rule solution. *IEEE. Trans. Inform. Theory*, IT–27(1):132–136, 1981.

[43] Arthur Nadas. Hidden Markov chains, the forward-backward algorithm, and initial statistics. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP–31(2):504–506, 1983.

[44] G. E. Noether. On a theorem of Pitman. *Ann. Math. Stat.*, 26:64–68, 1955.

[45] M. Peligrad. Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables. In *Dependence in Probability and Statistics.* Birkhauser, 1985.

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

[47] H.Vincent Poor and John B. Thomas. Memoryless discrete-time detection of a constant signal in m-dependent noise. *IEEE Trans. Inform. Theory*, IT–25(1):54–61, 1979.

[48] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[49] Adrian E. Raftery. A continuous multivariate exponential distribution. *Commun. Statist.-Theor. Meth.*, 13(8):947–965, 1984.

[50] H. L. Royden. *Real Analysis.* MacMillan, New York, second edition, 1968.

[51] John S. Sadowsky. A maximum variance model for robust detection and estimation with dependent data. *IEEE Trans. Inform. Theory*, IT–32:220–226, 1986.

[52] John S. Sadowsky and James A. Bucklew. A nonlocal approach for asymptotic memoryless detection theory. *IEEE Trans. Inform. Theory*, IT–32:115–120, 1986.

[53] Douglas W. Sauder and Evaggelos A. Geraniotis. Optimal and robust memoryless discrimination from dependent observations. *IEEE Trans. Inform. Theory*, IT–37:73–91, 1991.

[54] Shelupsky. An introduction to spherical coordinates. *American Mathematical Monthly*, 69(7):644–646, 1962.

[55] John Shore. On a relation between maximum likelihood classification and minimum relative-entropy classification. *IEEE Trans. Inform. Theory*, IT–30(6):851–854, 1984.

[56] John E. Shore and Rodney W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inform. Theory*, IT–26(1):26–37, 1980.

[57] H. L. Van Trees. *Detection, Estimation, and Modulation Theory—Part I.* Wiley, New York, 1968.