Detection, Estimation, and Modulation Theory

PART I | HARRY L. VAN TREES

Detection, Estimation, and Modulation Theory

PART I Detection, Estimation, and Linear Modulation Theory

HARRY L. VAN TREES
Massachusetts Institute of Technology

John Wiley and Sons, Inc. New York · London · Sydney

Library of Congress Catalog Card Number: 67–23331

GB 471 89955X

Printed in the United States of America

etaries, including Mrs. Jarmila e Tortorici, typed sections of the

an outgrowth of my research t, and I shall be glad to send our a on a regular reciprocal basis. pported by Lincoln Laboratory groups directed by Dr. Herbert research at M.I.T. was partly lational Aeronautics and Space Research Laboratory of Electedged.

Harry L. Van Trees

ng a Problem in the Doctrine of

La Détermination ces Orbites des

ly Bodies Moving About the Sun in rk, 1963.

ion," *Proc. Cambridge Philos. Soc.*,

blem of the Most Efficient Tests of *. London*, A **231**, 289, (1933).

olation von Stationären Zufälligen 5, 1941.

moothing of Stationary Time Series, rk, 1949 (originally published as a

Signals and Noise, McGraw-Hill,

Contents

1

Introduction

In these two books, we shall study three areas of statistical theory, which we have labeled detection theory, estimation theory, and modulation theory. The goal is to develop these theories in a common mathematical framework and to demonstrate how they can be used to solve a wealth of practical problems in many diverse physical situations.

In this chapter we present three outlines of the material. The first is a topical outline in which we develop a qualitative understanding of the three areas by examining some typical problems of interest. The second is a logical outline in which we explore the various methods of attacking the problems. The third is a chronological outline in which we explain the structure of the books.

1.1 TOPICAL OUTLINE

An easy way to explain what is meant by detection theory is to examine several physical situations that lead to detection theory problems.

A simple digital communication system is shown in Fig. 1.1. The source puts out a binary digit every T seconds. Our object is to transmit this sequence of digits to some other location. The channel available for transmitting the sequence depends on the particular situation. Typically, it could be a telephone line, a radio link, or an acoustical channel. For

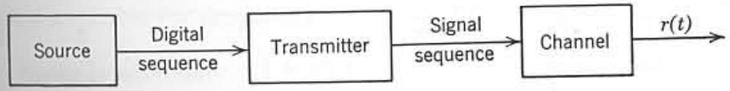


Fig. 1.1 Digital communication system.

purposes of illustration, we shall consider a radio link. In order to transmit the information, we must put it into a form suitable for propagating over the channel. A straightforward method would be to build a device that generates a sine wave,

$$s_1(t) = \sin \omega_1 t, \tag{1}$$

for T seconds if the source generated a "one" in the preceding interval, and a sine wave of a different frequency,

$$s_0(t) = \sin \omega_0 t, \tag{2}$$

for T seconds if the source generated a "zero" in the preceding interval. The frequencies are chosen so that the signals $s_0(t)$ and $s_1(t)$ will propagate over the particular radio link of concern. The output of the device is fed into an antenna and transmitted over the channel. Typical source and transmitted signal sequences are shown in Fig. 1.2. In the simplest kind of channel the signal sequence arrives at the receiving antenna attenuated but essentially undistorted. To process the received signal we pass it through the antenna and some stages of rf-amplification, in the course of which a thermal noise $n(t)$ is added to the message sequence. Thus in any T-second interval we have available a waveform $r(t)$ in which

$$r(t) = s_1(t) + n(t), \qquad 0 \le t \le T, \tag{3}$$

if $s_1(t)$ was transmitted, and

$$r(t) = s_0(t) + n(t), \qquad 0 \le t \le T, \tag{4}$$

if $s_0(t)$ was transmitted. We are now faced with the problem of deciding which of the two possible signals was transmitted. We label the device that does this a decision device. It is simply a processor that observes $r(t)$ and guesses whether $s_1(t)$ or $s_0(t)$ was sent according to some set of rules. This is equivalent to guessing what the source output was in the preceding interval. We refer to designing and evaluating the processor as a detection theory problem. In this particular case the only possible source of error in making a decision is the additive noise. If it were not present, the input would be completely known and we could make decisions without errors. We denote this type of problem as the *known signal in noise problem.* It corresponds to the lowest level (i.e., simplest) of the detection problems of interest.

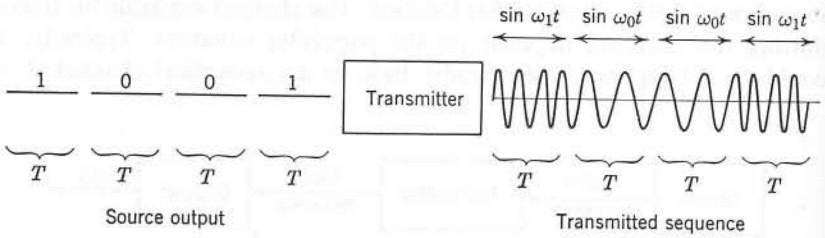


Fig. 1.2 Typical sequences.

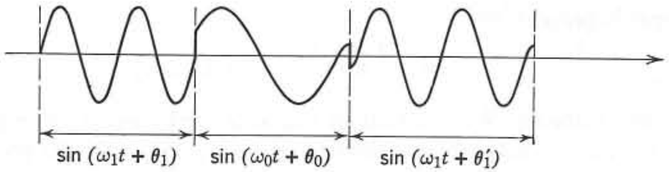

sin ($\omega_1 t + \theta_1$)
sin ($\omega_0 t + \theta_0$)
sin ($\omega_1 t + \theta_1'$)

Fig. 1.3 Sequence with phase shifts.

An example of the next level of detection problem is shown in Fig. 1.3. The oscillators used to generate $s_1(t)$ and $s_0(t)$ in the preceding example have a phase drift. Therefore in a particular T-second interval the received signal corresponding to a "one" is

$$r(t) = \sin(\omega_1 t + \theta_1) + n(t), \qquad 0 \le t \le T, \tag{5}$$

and the received signal corresponding to a "zero" is

$$r(t) = \sin(\omega_0 t + \theta_0) + n(t), \qquad 0 \le t \le T, \tag{6}$$

where θ_0 and θ_1 are unknown constant phase angles. Thus even in the absence of noise the input waveform is not completely known. In a practical system the receiver may include auxiliary equipment to measure the oscillator phase. If the phase varies slowly enough, we shall see that essentially perfect measurement is possible. If this is true, the problem is the same as above. However, if the measurement is not perfect, we must incorporate the signal uncertainty in our model.

A corresponding problem arises in the radar and sonar areas. A conventional radar transmits a pulse at some frequency ω_c with a rectangular envelope:

$$s_t(t) = \sin \omega_c t, \qquad 0 \le t \le T. \tag{7}$$

If a target is present, the pulse is reflected. Even the simplest target will introduce an attenuation and phase shift in the transmitted signal. Thus the signal available for processing in the interval of interest is

$$\begin{aligned} r(t) &= V_r \sin[\omega_c(t - \tau) + \theta_r] + n(t), && \tau \le t \le \tau + T, \\ &= n(t), && 0 \le t < \tau, \tau + T < t < \infty, \end{aligned} \tag{8}$$

if a target is present and

$$r(t) = n(t), \qquad 0 \leq t < \infty, \tag{9}$$

if a target is absent. We see that in the absence of noise the signal still contains three unknown quantities: V_r, the amplitude, θ_r, the phase, and τ, the round-trip travel time to the target.

These two examples represent the second level of detection problems. We classify them as *signal with unknown parameters in noise problems.*

Detection problems of a third level appear in several areas. In a passive sonar detection system the receiver listens for noise generated by enemy vessels. The engines, propellers, and other elements in the vessel generate acoustical signals that travel through the ocean to the hydrophones in the detection system. This composite signal can best be characterized as a sample function from a random process. In addition, the hydrophone generates self-noise and picks up sea noise. Thus a suitable model for the detection problem might be

$$r(t) = s_{\Omega}(t) + n(t) \tag{10}$$

if the target is present and

$$r(t) = n(t) \tag{11}$$

if it is not. In the absence of noise the signal is a sample function from a random process (indicated by the subscript Ω).

In the communications field a large number of systems employ channels in which randomness is inherent. Typical systems are tropospheric scatter links, orbiting dipole links, and chaff systems. A common technique is to transmit one of two signals separated in frequency. (We denote these frequencies as ω_1 and ω_0.) The resulting received signal is

$$r(t) = s_{\Omega_1}(t) + n(t) \tag{12}$$

if $s_1(t)$ was transmitted and

$$r(t) = s_{\Omega_0}(t) + n(t) \tag{13}$$

if $s_0(t)$ was transmitted. Here $s_{\Omega_1}(t)$ is a sample function from a random process centered at ω_1, and $s_{\Omega_0}(t)$ is a sample function from a random process centered at ω_0. These examples are characterized by the lack of any deterministic signal component. Any decision procedure that we design will have to be based on the difference in the statistical properties of the two random processes from which $s_{\Omega_0}(t)$ and $s_{\Omega_1}(t)$ are obtained. This is the third level of detection problem and is referred to as a *random signal in noise problem.*

In our examination of representative examples we have seen that detection theory problems are characterized by the fact that we must decide which of several alternatives is true. There were only two alternatives in the examples cited; therefore we refer to them as binary detection problems. Later we will encounter problems in which there are M alternatives available (the M-ary detection problem). Our hierarchy of detection problems is presented graphically in Fig. 1.4.

There is a parallel set of problems in the estimation theory area. A simple example is given in Fig. 1.5, in which the source puts out an analog message $a(t)$ (Fig. 1.5*a*). To transmit the message we first sample it every T seconds. Then, every T seconds we transmit a signal that contains

	Detection theory
Level 1. Known signals in noise	1. Synchronous digital communication 2. Pattern recognition problems
Level 2. Signals with unknown parameters in noise	1. Conventional pulsed radar or sonar, target detection 2. Target classification (orientation of target unknown) 3. Digital communication systems without phase reference 4. Digital communication over slowly-fading channels
Level 3. Random signals in noise	1. Digital communication over scatter link, orbiting dipole channel, or chaff link 2. Passive sonar 3. Seismic detection system 4. Radio astronomy (detection of noise sources)

Fig. 1.4 Detection theory hierarchy.

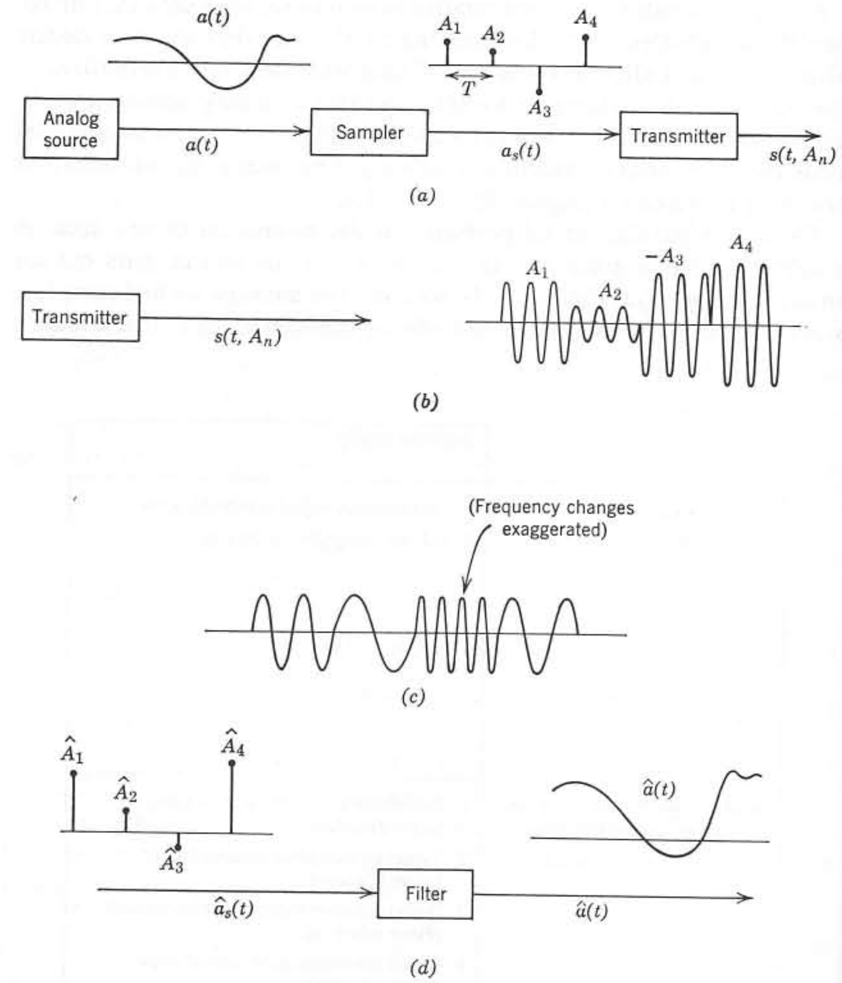


Fig. 1.5 **(*a*) Sampling an analog source; (*b*) pulse-amplitude modulation; (*c*) pulse-frequency modulation; (*d*) waveform reconstruction.**

a parameter which is uniquely related to the last sample value. In Fig. 1.5*b* the signal is a sinusoid whose amplitude depends on the last sample. Thus, if the sample at time nT is A_n, the signal in the interval $[nT, (n + 1)T]$ is

$$s(t, A_n) = A_n \sin \omega_c t, \qquad nT \le t \le (n + 1)T. \tag{14}$$

A system of this type is called a pulse amplitude modulation (PAM) system. In Fig. 1.5*c* the signal is a sinusoid whose frequency in the interval

differs from the reference frequency ω_c by an amount proportional to the preceding sample value,

$$s(t, A_n) = \sin (\omega_c t + A_n t), \qquad nT \le t \le (n + 1)T. \tag{15}$$

A system of this type is called a pulse frequency modulation (PFM) system. Once again there is additive noise. The received waveform, given that A_n was the sample value, is

$$r(t) = s(t, A_n) + n(t), \qquad nT \le t \le (n + 1)T. \tag{16}$$

During each interval the receiver tries to estimate A_n. We denote these estimates as $\hat{A}_n$. Over a period of time we obtain a sequence of estimates, as shown in Fig. 1.5*d*, which is passed into a device whose output is an estimate of the original message $a(t)$. If $a(t)$ is a band-limited signal, the device is just an ideal low-pass filter. For other cases it is more involved.

If, however, the parameters in this example were known and the noise were absent, the received signal would be completely known. We refer to problems in this category as *known signal in noise problems*. If we assume that the mapping from A_n to $s(t, A_n)$ in the transmitter has an inverse, we see that if the noise were not present we could determine A_n unambiguously. (Clearly, if we were allowed to design the transmitter, we should always choose a mapping with an inverse.) The *known signal in noise problem* is the first level of the estimation problem hierarchy.

Returning to the area of radar, we consider a somewhat different problem. We assume that we know a target is present but do not know its range or velocity. Then the received signal is

$$\begin{aligned} r(t) &= V_r \sin [(\omega_c + \omega_d)(t - \tau) + \theta_r] + n(t), \qquad && \tau \le t \le \tau + T, \\ &= n(t), && 0 \le t < \tau, \tau + T < t < \infty, \end{aligned} \tag{17}$$

where ω_d denotes a Doppler shift caused by the target's motion. We want to estimate τ and ω_d. Now, even if the noise were absent and τ and ω_d were known, the signal would still contain the unknown parameters V_r and θ_r. This is a typical second-level estimation problem. As in detection theory, we refer to problems in this category as *signal with unknown parameters in noise problems.*

At the third level the signal component is a random process whose statistical characteristics contain parameters we want to estimate. The received signal is of the form

$$r(t) = s_\Omega(t, A) + n(t), \tag{18}$$

where $s_\Omega(t, A)$ is a sample function from a random process. In a simple case it might be a stationary process with the narrow-band spectrum shown in Fig. 1.6. The shape of the spectrum is known but the center frequency

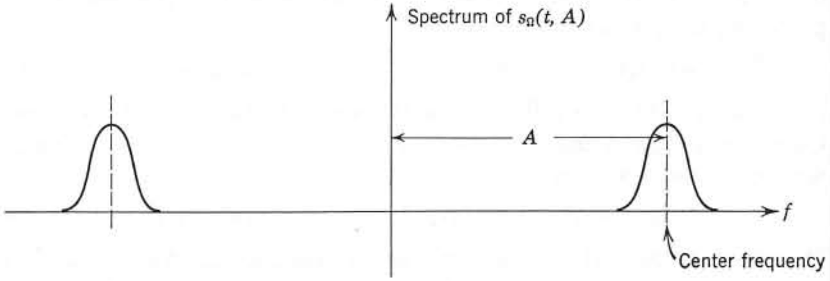


Fig. 1.6 Spectrum of random signal.

	Estimation Theory
Level 1. Known signals in noise	1. PAM, PFM, and PPM communication systems with phase synchronization 2. Inaccuracies in inertial systems (e.g., drift angle measurement)
Level 2. Signals with unknown parameters in noise	1. Range, velocity, or angle measurement in radar/sonar problems 2. Discrete time, continuous amplitude communication system (with unknown amplitude or phase in channel)
Level 3. Random signals in noise	1. Power spectrum parameter estimation 2. Range or Doppler spread target parameters in radar/sonar problem 3. Velocity measurement in radio astronomy 4. Target parameter estimation: passive sonar 5. Ground mapping radars

Fig. 1.7 Estimation theory hierarchy.

is not. The receiver must observe $r(t)$ and, using the statistical properties of $s_\Omega(t, A)$ and $n(t)$, estimate the value of A. This particular example could arise in either radio astronomy or passive sonar. The general class of problem in which the signal containing the parameters is a sample function from a random process is referred to as the *random signal in noise problem*. The hierarchy of estimation theory problems is shown in Fig. 1.7.

We note that there appears to be considerable parallelism in the detection and estimation theory problems. We shall frequently exploit these parallels to reduce the work, but there is a basic difference that should be emphasized. In binary detection the receiver is either "right" or "wrong." In the estimation of a continuous parameter the receiver will seldom be exactly right, but it can try to be close most of the time. This difference will be reflected in the manner in which we judge system performance.

The third area of interest is frequently referred to as modulation theory. We shall see shortly that this term is too narrow for the actual problems. Once again a simple example is useful. In Fig. 1.8 we show an analog message source whose output might typically be music or speech. To convey the message over the channel, we transform it by using a modulation scheme to get it into a form suitable for propagation. The transmitted signal is a continuous waveform that depends on $a(t)$ in some deterministic manner. In Fig. 1.8 it is an amplitude modulated waveform:

$$s[t, a(t)] = [1 + ma(t)] \sin(\omega_c t). \qquad (19)$$

(This is conventional double-sideband AM with modulation index m.) In Fig. 1.8*c* the transmitted signal is a frequency modulated (FM) waveform:

$$s[t, a(t)] = \sin\left[\omega_c t + \int_{-\infty}^{t} a(u)\, du\right]. \qquad (20)$$

When noise is added the received signal is

$$r(t) = s[t, a(t)] + n(t). \qquad (21)$$

Now the receiver must observe $r(t)$ and put out a continuous estimate of the message $a(t)$, as shown in Fig. 1.8. This particular example is a first-level modulation problem, for if $n(t)$ were absent and $a(t)$ were known the received signal would be completely known. Once again we describe it as a *known signal in noise problem*.

Another type of physical situation in which we want to estimate a continuous function is shown in Fig. 1.9. The channel is a time-invariant linear system whose impulse response $h(\tau)$ is unknown. To estimate the impulse response we transmit a known signal $x(t)$. The received signal is

$$r(t) = \int_0^\infty h(\tau)\, x(t - \tau)\, d\tau + n(t). \qquad (22)$$

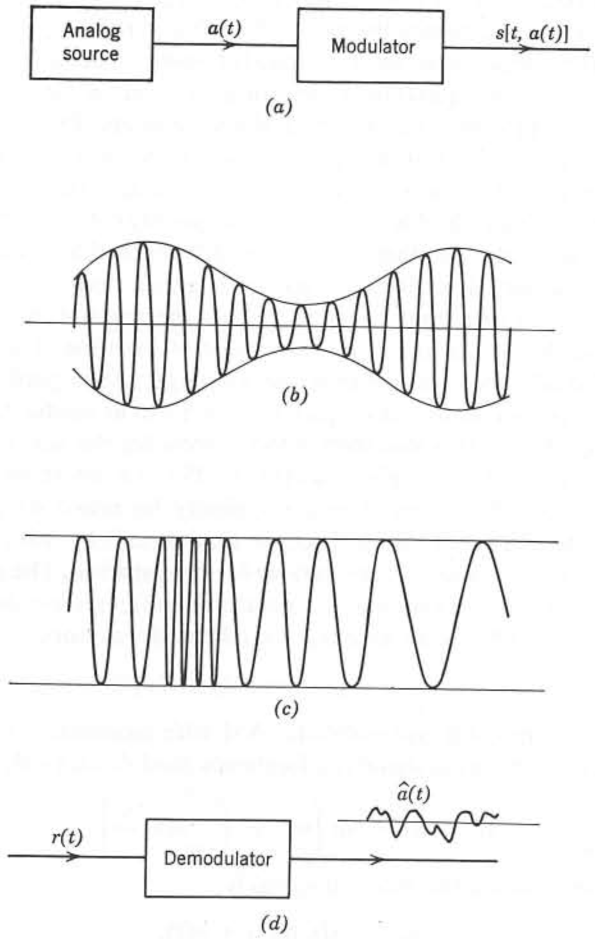


Fig. 1.8 A modulation theory example: (*a*) analog transmission system; (*b*) amplitude modulated signal; (*c*) frequency modulated signal; (*d*) demodulator.

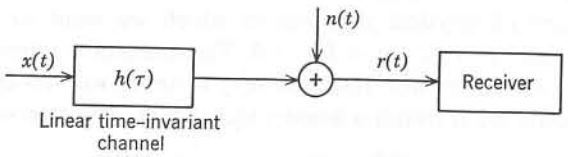


Fig. 1.9 Channel measurement.

The receiver observes $r(t)$ and tries to estimate $h(\tau)$. This particular example could best be described as a continuous estimation problem. Many other problems of interest in which we shall try to estimate a continuous waveform will be encountered. For convenience, we shall use the term *modulation theory* for this category, even though the term continuous waveform estimation might be more descriptive.

The other levels of the modulation theory problem follow by direct analogy. In the amplitude modulation system shown in Fig. 1.8*b* the receiver frequently does not know the phase angle of the carrier. In this case a suitable model is

$$r(t) = (1 + ma(t)) \sin (\omega_c t + \theta) + n(t), \tag{23}$$

	Modulation Theory (Continuous waveform estimation)
1. Known signals in noise	1. Conventional communication systems such as AM (DSB-AM, SSB), FM, and PM with phase synchronization 2. Optimum filter theory 3. Optimum feedback systems 4. Channel measurement 5. Orbital estimation for satellites 6. Signal estimation in seismic and sonar classification systems 7. Synchronization in digital systems
2. Signals with unknown parameters in noise	1. Conventional communication systems without phase synchronization 2. Estimation of channel characteristics when phase of input signal is unknown
3. Random signals in noise	1. Analog communication over randomly varying channels 2. Estimation of statistics of time-varying processes 3. Estimation of plant characteristics

Fig. 1.10 Modulation theory hierarchy.

where θ is an unknown parameter. This is an example of a *signal with unknown parameter problem* in the modulation theory area.

A simple example of a third-level problem (*random signal in noise*) is one in which we transmit a frequency-modulated signal over a radio link whose gain and phase characteristics are time-varying. We shall find that if we transmit the signal in (20) over this channel the received waveform will be

$$r(t) = V(t)\sin\left[\omega_c t + \int_{-\infty}^{t} a(u)\,du + \theta(t)\right] + n(t), \qquad (24)$$

where $V(t)$ and $\theta(t)$ are sample functions from random processes. Thus, even if $a(u)$ were known and the noise $n(t)$ were absent, the received signal would still be a random process. An over-all outline of the problems of interest to us appears in Fig. 1.10. Additional examples included in the table to indicate the breadth of the problems that fit into the outline are discussed in more detail in the text.

Now that we have outlined the areas of interest it is appropriate to determine how to go about solving them.

1.2 POSSIBLE APPROACHES

From the examples we have discussed it is obvious that an inherent feature of all the problems is randomness of source, channel, or noise (often all three). Thus our approach must be statistical in nature. Even assuming that we are using a statistical model, there are many different ways to approach the problem. We can divide the possible approaches into two categories, which we denote as "structured" and "nonstructured." Some simple examples will illustrate what we mean by a structured approach.

Example 1. The input to a linear time-invariant system is $r(t)$:

$$\begin{aligned} r(t) &= s(t) + w(t) \qquad 0 \le t \le T, \\ &= 0, \qquad\qquad\quad\ \text{elsewhere.} \end{aligned} \qquad (25)$$

The impulse response of the system is $h(\tau)$. The signal $s(t)$ is a known function with energy E_s,

$$E_s = \int_0^T s^2(t)\,dt, \qquad (26)$$

and $w(t)$ is a sample function from a zero-mean random process with a covariance function:

$$K_w(t, u) = \frac{N_0}{2}\,\delta(t-u). \qquad (27)$$

We are concerned with the output of the system at time T. The output due to the signal is a deterministic quantity:

$$s_o(T) = \int_0^T h(\tau)\,s(T-\tau)\,d\tau. \qquad (28)$$

The output due to the noise is a random variable:

$$n_o(T) = \int_0^T h(\tau)\,n(T-\tau)\,d\tau. \qquad (29)$$

We can define the output signal-to-noise ratio at time T as

$$\frac{S}{N} \triangleq \frac{s_o^2(T)}{E[n_o^2(T)]}, \qquad (30)$$

where $E(\cdot)$ denotes expectation.

Substituting (28) and (29) into (30), we obtain

$$\frac{S}{N} = \frac{\left[\int_0^T h(\tau)\,s(T-\tau)\,d\tau\right]^2}{E\left[\iint_0^T h(\tau)\,h(u)\,n(T-\tau)\,n(T-u)\,d\tau\,du\right]}. \qquad (31)$$

By bringing the expectation inside the integral, using (27), and performing the integration with respect to u, we have

$$\frac{S}{N} = \frac{\left[\int_0^T h(\tau)\,s(T-\tau)\,d\tau\right]^2}{N_0/2\int_0^T h^2(\tau)\,d\tau}. \qquad (32)$$

The problem of interest is to choose $h(\tau)$ to maximize the signal-to-noise ratio. The solution follows easily, but it is not important for our present discussion. (See Problem 3.3.1.)

This example illustrates the three essential features of the structured approach to a statistical optimization problem:

Structure. The processor was required to be a linear time-invariant filter. We wanted to choose the best system in this class. Systems that were not in this class (e.g., nonlinear or time-varying) were not allowed.

Criterion. In this case we wanted to maximize a quantity that we called the signal-to-noise ratio.

Information. To write the expression for S/N we had to know the signal shape and the covariance function of the noise process.

If we knew more about the process (e.g., its first-order probability density), we could not use it, and if we knew less, we could not solve the problem. Clearly, if we changed the criterion, the information required might be different. For example, to maximize x

$$x = \frac{s_o^4(T)}{E[n_o^4(T)]}, \qquad (33)$$

the covariance function of the noise process would not be adequate. Alternatively, if we changed the structure, the information required might change. Thus the three ideas of structure, criterion, and information are closely related. It is important to emphasize that the structured approach does not imply a linear system, as illustrated by Example 2.

Example 2. The input to the nonlinear no-memory device shown in Fig. 1.11 is $r(t)$, where

$$r(t) = s(t) + n(t), \qquad -\infty < t < \infty. \tag{34}$$

At any time t, $s(t)$ is the value of a random variable s with known probability density $p_s(S)$. Similarly, $n(t)$ is the value of a statistically independent random variable n with known density $p_n(N)$. The output of the device is $y(t)$, where

$$y(t) = a_0 + a_1[r(t)] + a_2[r(t)]^2 \tag{35}$$

is a quadratic no-memory function of $r(t)$. [The adjective no-memory emphasizes that the value of $y(t_0)$ depends *only* on $r(t_0)$.] We want to choose the coefficients a_0, a_1, and a_2 so that $y(t)$ is the minimum mean-square error estimate of $s(t)$. The mean-square error is

$$\begin{aligned}\xi(t) &\triangleq E\{[y(t) - s(t)^2]\} \\ &= E(\{a_0 + a_1[r(t)] + a_2[r^2(t)] - s(t)\}^2)\end{aligned} \tag{36}$$

and a_0, a_1, and a_2 are chosen to minimize $\xi(t)$. The solution to this particular problem is given in Chapter 3.

The technique for solving structured problems is conceptually straightforward. We allow the structure to vary within the allowed class and choose the particular system that maximizes (or minimizes) the criterion of interest.

An obvious advantage to the structured approach is that it usually requires only a partial characterization of the processes. This is important because, in practice, we must measure or calculate the process properties needed.

An obvious disadvantage is that it is often impossible to tell if the structure chosen is correct. In Example 1 a simple nonlinear system might be far superior to the best linear system. Similarly, in Example 2 some other nonlinear system might be far superior to the quadratic system. Once a class of structure is chosen we are committed. A number of trivial examples demonstrate the effect of choosing the wrong structure. We shall encounter an important practical example when we study frequency modulation in Chapter II-2.

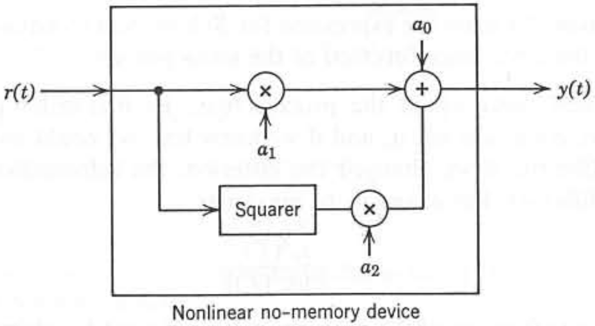


Fig. 1.11 A structured nonlinear device.

At first glance it appears that one way to get around the problem of choosing the proper strucutre is to let the structure be an arbitrary nonlinear time-varying system. In other words, the class of structure is chosen to be so large that every possible system will be included in it. The difficulty is that there is no convenient tool, such as the convolution integral, to express the output of a nonlinear system in terms of its input. This means that there is no convenient way to investigate all possible systems by using a structured approach.

The alternative to the structured approach is a nonstructured approach. Here we refuse to make any a priori guesses about what structure the processor should have. We establish a criterion, solve the problem, and implement whatever processing procedure is indicated.

A simple example of the nonstructured approach can be obtained by modifying Example 2. Instead of assigning characteristics to the device, we denote the estimate by $y(t)$. Letting

$$\xi(t) \triangleq E\{[y(t) - s(t)]^2\}, \tag{37}$$

we solve for the $y(t)$ that is obtained from $r(t)$ in *any* manner to minimize ξ. The obvious advantage is that if we can solve the problem we know that our answer, is *with respect to the chosen criterion*, the best processor of all possible processors. The obvious disadvantage is that we must completely characterize all the signals, channels, and noises that enter into the problem. Fortunately, it turns out that there are a large number of problems of practical importance in which this complete characterization is possible. Throughout both books we shall emphasize the nonstructured approach.

Our discussion up to this point has developed the topical and logical basis of these books. We now discuss the actual organization.

1.3 ORGANIZATION

The material covered in this book and Volume II can be divided into five parts. The first can be labeled *Background* and consists of Chapters 2 and 3. In Chapter 2 we develop in detail a topic that we call Classical Detection and Estimation Theory. Here we deal with problems in which the observations are sets of random variables instead of random waveforms. The theory needed to solve problems of this type has been studied by statisticians for many years. We therefore use the adjective classical to describe it. The purpose of the chapter is twofold: first, to derive all the basic statistical results we need in the remainder of the chapters; second, to provide a general background in detection and estimation theory that can be extended into various areas that we do not discuss in detail. To accomplish the second purpose we keep the discussion as general as possible. We consider in detail the binary and M-ary hypothesis testing problem, the problem of estimating random and nonrandom variables, and the composite hypothesis testing problem. Two more specialized topics, the general Gaussian problem and performance bounds on binary tests, are developed as background for specific problems we shall encounter later.

The next step is to bridge the gap between the classical case and the waveform problems discussed in Section 1.1. Chapter 3 develops the necessary techniques. The key to the transition is a suitable method for characterizing random processes. When the observation interval is finite, the most useful characterization is by a series expansion of the random process which is a generalization of the conventional Fourier series. When the observation interval is infinite, a transform characterization, which is a generalization of the usual Fourier transform, is needed. In the process of developing these characterizations, we encounter integral equations and we digress briefly to develop methods of solution. Just as in Chapter 2, our discussion is general and provides background for other areas of application.

With these two chapters in the first part as background, we are prepared to work our way through the hierarchy of problems outlined in Figs. 1.4, 1.7, and 1.10. The second part of the book (Chapter 4) can be labeled *Elementary Detection and Estimation Theory*. Here we develop the first two levels described in Section 1.1. (This material corresponds to the upper two levels in Figs. 1.4 and 1.7.) We begin by looking at the simple binary digital communication system described in Fig. 1.1 and then proceed to more complicated problems in the communications, radar, and sonar area involving M-ary communication, random phase channels, random amplitude and phase channels, and colored noise interference. By exploiting the parallel nature of the estimation problem, results are obtained easily for the estimation problem outlined in Fig. 1.5 and other more complex systems. The extension of the results to include the multiple channel (e.g., frequency diversity systems or arrays) and multiple parameter (e.g., range and Doppler) problems completes our discussion. The results in this chapter are fundamental to the understanding of modern communication and radar/sonar systems.

The third part, which can be labeled *Modulation Theory or Continuous Estimation Theory*, consists of Chapters 5 and 6 and Chapter 2 of Volume II. In Chapter 5 we formulate a quantitative model for the first two levels of the continuous waveform estimation problem and derive a set of integral equations whose solution is the optimum estimate of the message. We also derive equations that give bounds on the performance of the estimators. In order to study solution techniques, we divide the estimation problem into two categories, linear and nonlinear.

In Chapter 6 we study linear estimation problems in detail. In the first section of the chapter we discuss the relationships between various criteria, process characteristics, and the structure of the processor. In the next section we discuss the special case in which the processes are stationary and the infinite past is available. This case, the Wiener problem, leads to straightforward solution techniques. The original work of Wiener is extended to obtain some important closed-form error expressions. In the next section we discuss the case in which the processes can be characterized by using state-variable techniques. This case, the Kalman-Bucy problem, enables us to deal with nonstationary, finite-interval problems and adds considerable insight to the results of the preceding section.

The material in Chapters 1 through 6 has two characteristics:

1. In almost all cases we can obtain *explicit*, *exact* solutions to the problems that we formulate.

2. Most of the topics discussed are of such fundamental interest that everyone concerned with the statistical design of communication, radar, or sonar systems should be familiar with them.

As soon as we try to solve the nonlinear estimation problem, we see a sharp departure. To obtain useful results we must resort to approximate solution techniques. To decide what approximations are valid, however, we must consider specific nonlinear modulation systems. Thus the precise quantitative results are only applicable to the specific system. In view of this departure, we pause briefly in our logical development and summarize our results in Chapter 7.

After a brief introduction we return to the nonlinear modulation problem in Chapter 2 of Volume II and consider angle modulation systems in great detail. After an approximation to the optimum processor is developed, its performance and possible design modification are analyzed both theoretically and experimentally. More advanced techniques from Markov process theory and information theory are used to obtain significant results.

In the fourth part we revisit the problems of detection, estimation, and modulation theory at the third level of the hierarchy described in Section 1.1. Looking at the bottom boxes in Figs. 1.4, 1.7, and 1.10, we see that this is the *Random Signals in Noise* problem. Chapter II-3 studies it in detail. We find that the linear processors developed in Chapter I-6 play a fundamental role in the random signal problem. This result, coupled with the corresponding result in Chapter II-2, emphasizes the fundamental importance of the results in Chapter I-6. They also illustrate the inherent unity of the various problems. Specific topics such as power-spectrum parameter estimation and analog transmission over time-varying channels are also developed.

The fifth part is labeled *Applications* and includes Chapters II-4 and II-5. Throughout the two books we emphasize applications of the theory to models of practical problems. In most of them the relation of the actual physical situation can be explained in a page or two. The fifth part deals with physical situations in which developing the model from the physical situation is a central issue. Chapter II-4 studies the radar/sonar problem in depth. It builds up a set of target and channel models, starting with slowly fluctuating point targets and culminating in deep targets that fluctuate at arbitrary rates. This set of models enables us to study the signal design problem for radar and sonar, the resolution problem in mapping radars, the effect of reverberation on sonar-system performance, estimation of parameters of spread targets, communication over spread channels, and other important problems.

In Chapter II-5 we study various multidimensional problems such as multiplex communication systems and multivariable processing problems encountered in continuous receiving apertures and optical systems. The primary emphasis in the chapter is on optimum array processing in sonar (or seismic) systems. Both active and passive sonar systems are discussed; specific processor configurations are developed and their performance is analyzed.

Finally, in Chapter II-6 we summarize some of the more important results, mention some related topics that have been omitted, and suggest areas of future research.

2

Classical Detection and Estimation Theory

2.1 INTRODUCTION

In this chapter we develop in detail the basic ideas of classical detection and estimation theory. The first step is to define the various terms.

The basic components of a simple decision-theory problem are shown in Fig. 2.1. The first is a *source* that generates an output. In the simplest case this output is one of two choices. We refer to them as hypotheses and label them H_0 and H_1 in the two-choice case. More generally, the output might be one of M hypotheses, which we label $H_0, H_1, \ldots, H_{M-1}$. Some typical source mechanisms are the following:

1. A digital communication system transmits information by sending ones and zeros. When "one" is sent, we call it H_1, and when "zero" is sent, we call it H_0.

2. In a radar system we look at a particular range and azimuth and try

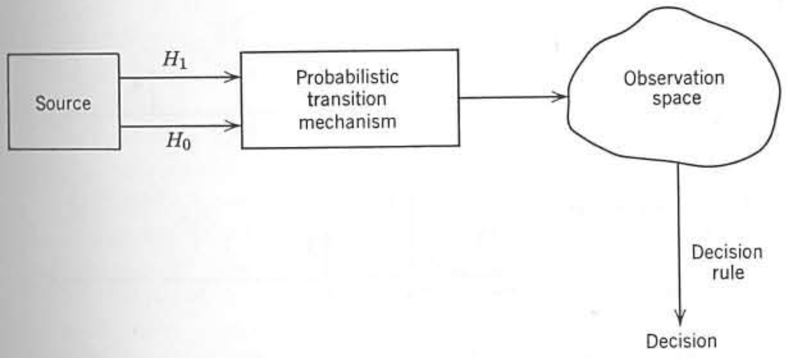


Fig. 2.1 Components of a decision theory problem.

to decide whether a target is present; H_1 corresponds to the presence of a target and H_0 corresponds to no target.

3. In a medical diagnosis problem we examine an electrocardiogram. Here H_1 could correspond to the patient having had a heart attack and H_0 to the absence of one.

4. In a speaker classification problem we know the speaker is German, British, or American and either male or female. There are six possible hypotheses.

In the cases of interest to us we do not know which hypothesis is true.

The second component of the problem is a *probabilistic transition mechanism*; the third is an *observation space*. The transition mechanism can be viewed as a device that knows which hypothesis is true. Based on this knowledge, it generates a point in the observation space according to some probability law.

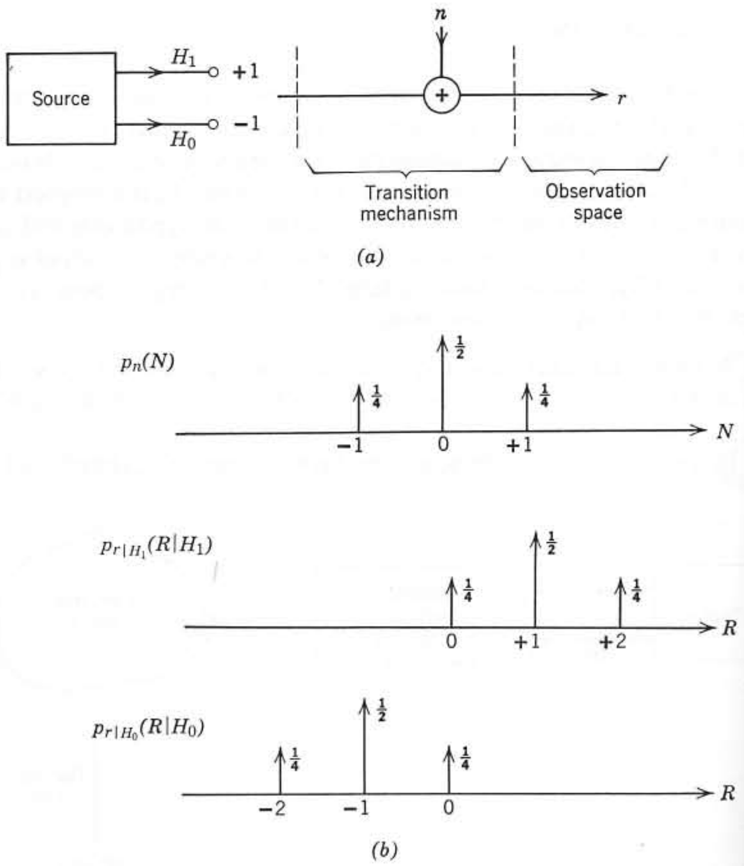


Fig. 2.2 A simple decision problem: (*a*) model; (*b*) probability densities.

A simple example to illustrate these ideas is given in Fig. 2.2. When H_1 is true, the source generates $+1$. When H_0 is true, the source generates -1. An independent discrete random variable n whose probability density is shown in Fig. 2.2*b* is added to the source output. The sum of the source output and n is the observed variable r.

Under the two hypotheses, we have

$$\begin{aligned} H_1 &: r = 1 + n, \\ H_0 &: r = -1 + n. \end{aligned} \tag{1}$$

The probability densities of r on the two hypotheses are shown in Fig. 2.2*b*. The observation space is one-dimensional, for any output can be plotted on a line.

A related example is shown in Fig. 2.3*a* in which the source generates two numbers in sequence. A random variable n_1 is added to the first number and an independent random variable n_2 is added to the second. Thus

$$\begin{aligned} H_1 &: r_1 = 1 + n_1 \\ &\quad\, r_2 = 1 + n_2, \\ H_0 &: r_1 = -1 + n_1 \\ &\quad\, r_2 = -1 + n_2. \end{aligned} \tag{2}$$

The joint probability density of r_1 and r_2 when H_1 is true is shown in Fig. 2.3*b*. The observation space is two-dimensional and any observation can be represented as a point in a plane.

In this chapter we confine our discussion to problems in which the observation space is finite-dimensional. In other words, the observations consist of a set of N numbers and can be represented as a point in an N-dimensional space. This is the class of problem that statisticians have treated for many years. For this reason we refer to it as the *classical* decision problem.

The fourth component of the detection problem is a *decision* rule. After observing the outcome in the observation space we shall guess which hypothesis was true, and to accomplish this we develop a decision rule that assigns each point to one of the hypotheses. Suitable choices for decision rules will depend on several factors which we discuss in detail later. Our study will demonstrate how these four components fit together to form the total decision (or hypothesis-testing) problem.

The classical estimation problem is closely related to the detection problem. We describe it in detail later.

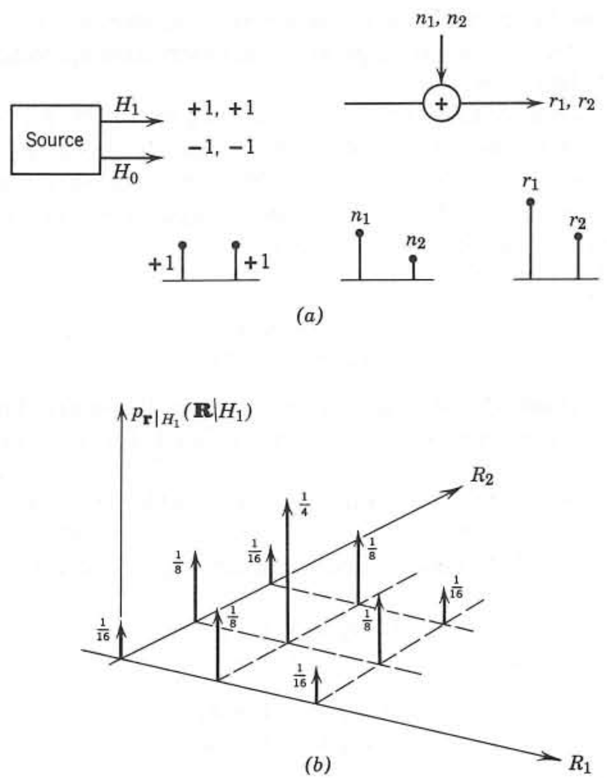


Fig. 2.3 A two-dimensional problem: (*a*) model; (*b*) probability density.

Organization. This chapter is organized in the following manner. In Section 2.2 we study the binary hypothesis testing problem. Then in Section 2.3 we extend the results to the case of M hypotheses. In Section 2.4 classical estimation theory is developed.

The problems that we encounter in Sections 2.2 and 2.3 are characterized by the property that each source output corresponds to a different hypothesis. In Section 2.5 we shall examine the composite hypothesis testing problem. Here a number of source outputs are lumped together to form a single hypothesis.

All of the developments through Section 2.5 deal with arbitrary probability transition mechanisms. In Section 2.6 we consider in detail a special class of problems that will be useful in the sequel. We refer to it as the general Gaussian class.

In many cases of practical importance we can develop the "optimum" decision rule according to certain criteria but cannot evaluate how well the test will work. In Section 2.7 we develop bounds and approximate expressions for the performance that will be necessary for some of the later chapters.

Finally, in Section 2.8 we summarize our results and indicate some of the topics that we have omitted.

2.2 SIMPLE BINARY HYPOTHESIS TESTS

As a starting point we consider the decision problem in which each of two source outputs corresponds to a hypothesis. Each hypothesis maps into a point in the observation space. We assume that the observation space corresponds to a set of N observations: $r_1, r_2, r_3, \ldots, r_N$. Thus each set can be thought of as a point in an N-dimensional space and can be denoted by a vector $\mathbf{r}$:

$$\mathbf{r} \triangleq \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix} \tag{3}$$

The probabilistic transition mechanism generates points in accord with the two known conditional probability densities $p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)$ and $p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)$. The object is to use this information to develop a suitable decision rule. To do this we must look at various criteria for making decisions.

2.2.1 Decision Criteria

In the binary hypothesis problem we know that either H_0 or H_1 is true. We shall confine our discussion to decision rules that are required to make a choice. (An alternative procedure would be to allow decision rules with three outputs (a) H_0 true, (b) H_1 true, (c) don't know.) Thus each time the experiment is conducted one of four things can happen:

1. H_0 true; choose H_0.
2. H_0 true; choose H_1.
3. H_1 true; choose H_1.
4. H_1 true; choose H_0.

The first and third alternatives correspond to correct choices. The second and fourth alternatives correspond to errors. The purpose of a decision criterion is to attach some relative importance to the four possible courses of action. It might be expected that the method of processing the received data (**r**) would depend on the decision criterion we select. In this section we show that for the two criteria of most interest, the Bayes and the Neyman–Pearson, the operations on **r** are identical.

Bayes Criterion. A Bayes test is based on two assumptions. The first is that the source outputs are governed by probability assignments, which are denoted by P_1 and P_0, respectively, and called the a priori probabilities. These probabilities represent the observer's information about the source before the experiment is conducted. The second assumption is that a cost is assigned to each possible course of action. We denote the cost for the four courses of action as C_{00}, C_{10}, C_{11}, C_{01}, respectively. The first subscript indicates the hypothesis chosen and the second, the hypothesis that was true. Each time the experiment is conducted a certain cost will be incurred. We should like to design our decision rule so that *on the average* the cost will be as small as possible. To do this we first write an expression for the expected value of the cost. We see that there are two probabilities that we must average over; the a priori probability and the probability that a particular course of action will be taken. Denoting the expected value of the cost as the risk $\mathcal{R}$, we have:

$$\begin{aligned}\mathcal{R} = {} & C_{00}P_0 \Pr(\text{say } H_0|H_0 \text{ is true}) \\ & + C_{10}P_0 \Pr(\text{say } H_1|H_0 \text{ is true}) \\ & + C_{11}P_1 \Pr(\text{say } H_1|H_1 \text{ is true}) \\ & + C_{01}P_1 \Pr(\text{say } H_0|H_1 \text{ is true}). \end{aligned} \tag{4}$$

Because we have assumed that the decision rule must say either H_1 or H_0, we can view it as a rule for dividing the total observation space Z into two parts, Z_0 and Z_1, as shown in Fig. 2.4. Whenever an observation falls in Z_0 we say H_0, and whenever an observation falls in Z_1 we say H_1.

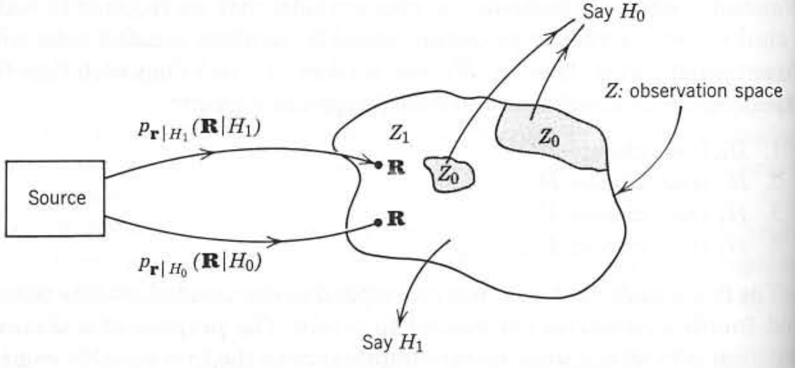


Fig. 2.4 Decision regions.

We can now write the expression for the risk in terms of the transition probabilities and the decision regions:

$$\begin{aligned}\mathcal{R} = {} & C_{00}P_0 \int_{Z_0} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} \\ & + C_{10}P_0 \int_{Z_1} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} \\ & + C_{11}P_1 \int_{Z_1} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R} \\ & + C_{01}P_1 \int_{Z_0} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R}. \end{aligned} \tag{5}$$

For an N-dimensional observation space the integrals in (5) are N-fold integrals.

We shall assume throughout our work that the cost of a wrong decision is higher than the cost of a correct decision. In other words,

$$\begin{aligned} C_{10} &> C_{00}, \\ C_{01} &> C_{11}. \end{aligned} \tag{6}$$

Now, to find the Bayes test we must choose the decision regions Z_0 and Z_1 in such a manner that the risk will be minimized. Because we require that a decision be made, this means that we must assign each point $\mathbf{R}$ in the observation space Z to Z_0 or Z_1.

Thus

$$Z = Z_0 + Z_1 \triangleq Z_0 \cup Z_1. \tag{7}$$

Rewriting (5), we have

$$\begin{aligned}\mathcal{R} = {} & P_0C_{00} \int_{Z_0} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} + P_0C_{10} \int_{Z-Z_0} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} \\ & + P_1C_{01} \int_{Z_0} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R} + P_1C_{11} \int_{Z-Z_0} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R}. \end{aligned} \tag{8}$$

Observing that

$$\int_Z p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} = \int_Z p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R} = 1, \tag{9}$$

(8) reduces to

$$\begin{aligned}\mathcal{R} = {} & P_0C_{10} + P_1C_{11} \\ & + \int_{Z_0} \{[P_1(C_{01} - C_{11})p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)] \\ & - [P_0(C_{10} - C_{00})p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)]\}\, d\mathbf{R}. \end{aligned} \tag{10}$$

The first two terms represent the fixed cost. The integral represents the cost controlled by those points $\mathbf{R}$ that we assign to Z_0. The assumption in (6) implies that the two terms inside the brackets are positive. Therefore all values of $\mathbf{R}$ where the second term is larger than the first should be included in Z_0 because they contribute a negative amount to the integral. Similarly, all values of $\mathbf{R}$ where the first term is larger than the second should be excluded from Z_0 (assigned to Z_1) because they would contribute a positive amount to the integral. Values of $\mathbf{R}$ where the two terms are equal have no effect on the cost and may be assigned arbitrarily. We shall assume that these points are assigned to H_1 and ignore them in our subsequent discussion. Thus the decision regions are defined by the statement: If

$$P_1(C_{01} - C_{11})p_{\mathbf{r}|H_1}(\mathbf{R}|H_1) \geq P_0(C_{10} - C_{00})p_{\mathbf{r}|H_0}(\mathbf{R}|H_0), \tag{11}$$

assign $\mathbf{R}$ to Z_1 and consequently say that H_1 is true. Otherwise assign $\mathbf{R}$ to Z_0 and say H_0 is true.

Alternately, we may write

$$\frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})}. \tag{12}$$

The quantity on the left is called the *likelihood ratio* and denoted by $\Lambda(\mathbf{R})$

$$\boxed{\Lambda(\mathbf{R}) \triangleq \frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)}.} \tag{13}$$

Because it is the ratio of two functions of a random variable, it is a random variable. We see that regardless of the dimensionality of $\mathbf{R}$, $\Lambda(\mathbf{R})$ is a one-dimensional variable.

The quantity on the right of (12) is the threshold of the test and is denoted by η:

$$\eta \triangleq \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})}. \tag{14}$$

Thus Bayes criterion leads us to a *likelihood ratio test* (LRT)

$$\Lambda(\mathbf{R}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{15}$$

We see that all the data processing is involved in computing $\Lambda(\mathbf{R})$ and is not affected by a priori probabilities or cost assignments. This invariance of the data processing is of considerable practical importance. Frequently the costs and a priori probabilities are merely educated guesses. The result in (15) enables us to build the entire processor and leave η as a variable threshold to accommodate changes in our estimates of a priori probabilities and costs.

Because the natural logarithm is a monotonic function, and both sides of (15) are positive, an equivalent test is

$$\ln \Lambda(\mathbf{R}) \underset{H_0}{\overset{H_1}{\gtrless}} \ln \eta. \tag{16}$$

Two forms of a processor to implement a likelihood ratio test are shown in Fig. 2.5.

Before proceeding to other criteria, we consider three simple examples.

Example 1. We assume that under H_1 the source output is a constant voltage m. Under H_0 the source output is zero. Before observation the voltage is corrupted by an additive noise. We sample the output waveform each second and obtain N samples. Each noise sample is a zero-mean Gaussian random variable n with variance σ^2. The noise samples at various instants are independent random variables and are independent of the source output. Looking at Fig. 2.6, we see that the observations under the two hypotheses are

$$\begin{aligned} H_1: r_i &= m + n_i \qquad i = 1, 2, \ldots, N, \\ H_0: r_i &= \phantom{m +{}} n_i \qquad i = 1, 2, \ldots, N, \end{aligned} \tag{17}$$

and

$$p_{n_i}(X) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{X^2}{2\sigma^2}\right), \tag{18}$$

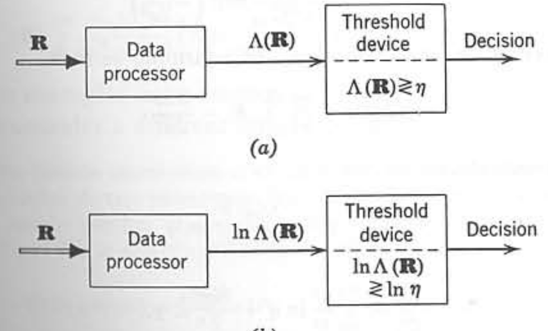
because the noise samples are Gaussian.

The probability density of r_i under each hypothesis follows easily:

$$p_{r_i|H_1}(R_i|H_1) = p_{n_i}(R_i - m) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(R_i - m)^2}{2\sigma^2}\right) \tag{19}$$

and

$$p_{r_i|H_0}(R_i|H_0) = p_{n_i}(R_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R_i^2}{2\sigma^2}\right). \tag{20}$$



(a)

(b)

Fig. 2.5 Likelihood ratio processors.

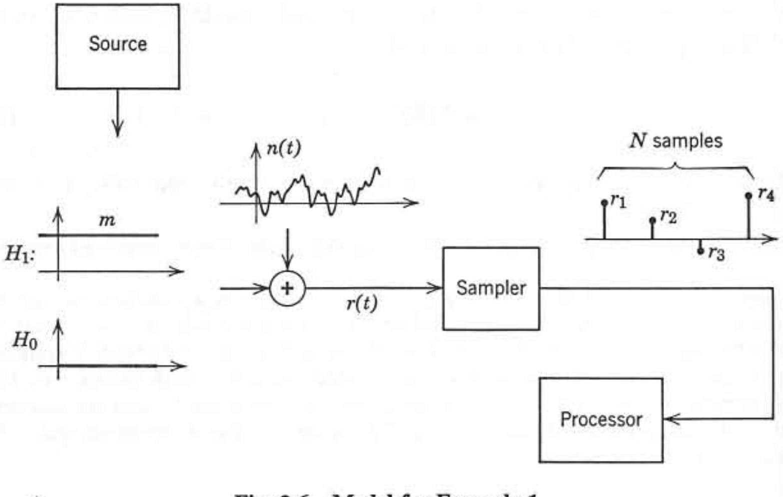


Fig. 2.6 Model for Example 1.

Because the n_i are statistically independent, the joint probability density of the r_i (or, equivalently, of the vector $\mathbf{r}$) is simply the product of the individual probability densities. Thus

$$p_{\mathbf{r}|H_1}(\mathbf{R}|H_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(R_i - m)^2}{2\sigma^2}\right), \qquad (21)$$

and

$$p_{\mathbf{r}|H_0}(\mathbf{R}|H_0) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R_i^2}{2\sigma^2}\right). \qquad (22)$$

Substituting into (13), we have

$$\Lambda(\mathbf{R}) = \frac{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(R_i - m)^2}{2\sigma^2}\right)}{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R_i^2}{2\sigma^2}\right)}. \qquad (23)$$

After canceling common terms and taking the logarithm, we have

$$\ln \Lambda(\mathbf{R}) = \frac{m}{\sigma^2} \sum_{i=1}^{N} R_i - \frac{Nm^2}{2\sigma^2}. \qquad (24)$$

Thus the likelihood ratio test is

$$\frac{m}{\sigma^2} \sum_{i=1}^{N} R_i - \frac{Nm^2}{2\sigma^2} \underset{H_0}{\overset{H_1}{\gtrless}} \ln \eta \qquad (25)$$

or, equivalently,

$$\sum_{i=1}^{N} R_i \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\sigma^2}{m} \ln \eta + \frac{Nm}{2} \triangleq \gamma. \qquad (26)$$

We see that the processor simply *adds* the observations and compares them with a threshold.

In this example the only way the data appear in the likelihood ratio test is in a sum. This is an example of a *sufficient statistic*, which we denote by $l(\mathbf{R})$ (or simply l when the argument is obvious). It is just a function of the received data which has the property that $\Lambda(\mathbf{R})$ can be written as a function of l. In other words, when making a decision, knowing the value of the sufficient statistic is just as good as knowing $\mathbf{R}$. In Example 1, l is a linear function of the R_i. A case in which this is not true is illustrated in Example 2.

Example 2. Several different physical situations lead to the mathematical model of interest in this example. The observations consist of a set of N values: $r_1, r_2, r_3, \ldots, r_N$. Under both hypotheses, the r_i are independent, identically distributed, zero-mean Gaussian random variables. Under H_1 each r_i has a variance σ_1^2. Under H_0 each r_i has a variance σ_0^2. Because the variables are independent, the joint density is simply the product of the individual densities. Therefore

$$p_{\mathbf{r}|H_1}(\mathbf{R}|H_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left(-\frac{R_i^2}{2\sigma_1^2}\right) \qquad (27)$$

and

$$p_{\mathbf{r}|H_0}(\mathbf{R}|H_0) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left(-\frac{R_i^2}{2\sigma_0^2}\right). \qquad (28)$$

Substituting (27) and (28) into (13) and taking the logarithm, we have

$$\frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right) \sum_{i=1}^{N} R_i^2 + N \ln \frac{\sigma_0}{\sigma_1} \underset{H_0}{\overset{H_1}{\gtrless}} \ln \eta. \qquad (29)$$

In this case the sufficient statistic is the sum of the squares of the observations

$$l(\mathbf{R}) = \sum_{i=1}^{N} R_i^2, \qquad (30)$$

and an equivalent test for $\sigma_1^2 > \sigma_0^2$ is

$$l(\mathbf{R}) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}\left(\ln \eta - N \ln \frac{\sigma_0}{\sigma_1}\right) \triangleq \gamma. \qquad (31)$$

For $\sigma_1^2 < \sigma_0^2$ the inequality is reversed because we are multiplying by a negative number:

$$l(\mathbf{R}) \underset{H_1}{\overset{H_0}{\gtrless}} \frac{2\sigma_0^2\sigma_1^2}{\sigma_0^2 - \sigma_1^2}\left(N \ln \frac{\sigma_0}{\sigma_1} - \ln \eta\right) \triangleq \gamma'; \qquad (\sigma_1^2 < \sigma_0^2). \qquad (32)$$

These two examples have emphasized Gaussian variables. In the next example we consider a different type of distribution.

Example 3. The Poisson distribution of events is encountered frequently as a model of shot noise and other diverse phenomena (e.g., [1] or [2]). Each time the experiment is conducted a certain number of events occur. Our observation is just this number which ranges from 0 to ∞ and obeys a Poisson distribution on both hypotheses; that is,

$$\Pr(n \text{ events}) = \frac{(m_i)^n}{n!} e^{-m_i}, \qquad n = 0, 1, 2 \ldots, i = 0, 1, \qquad (33)$$

where m_i is the parameter that specifies the average number of events:

$$E(n) = m_i. \qquad (34)$$

It is this parameter m_i that is different in the two hypotheses. Rewriting (33) to emphasize this point, we have for the two Poisson distributions

$$H_1\!:\!\Pr\,(n \text{ events}) = \frac{m_1^{\,n}}{n!}\,e^{-m_1}, \qquad n = 0, 1, 2, \ldots, \tag{35}$$

$$H_0\!:\!\Pr\,(n \text{ events}) = \frac{m_0^{\,n}}{n!}\,e^{-m_0}, \qquad n = 0, 1, 2, \ldots. \tag{36}$$

Then the likelihood ratio test is

$$\Lambda(n) = \left(\frac{m_1}{m_0}\right)^n \exp\left[-(m_1 - m_0)\right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{37}$$

or, equivalently,

$$\begin{aligned} n &\underset{H_0}{\overset{H_1}{\gtrless}} \frac{\ln \eta + m_1 - m_0}{\ln m_1 - \ln m_0}, &&\text{if } m_1 > m_0,\\ n &\underset{H_1}{\overset{H_0}{\gtrless}} \frac{\ln \eta + m_1 - m_0}{\ln m_1 - \ln m_0}, &&\text{if } m_0 > m_1. \end{aligned} \tag{38}$$

This example illustrates how the likelihood ratio test which we originally wrote in terms of probability densities can be simply adapted to accommodate observations that are discrete random variables. We now return to our general discussion of Bayes tests.

There are several special kinds of Bayes test which are frequently used and which should be mentioned explicitly.

If we assume that C_{00} and C_{11} are zero and $C_{01} = C_{10} = 1$, the expression for the risk in (8) reduces to

$$\mathcal{R} = P_0 \int_{Z_1} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} + P_1 \int_{Z_0} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R}. \tag{39}$$

We see that (39) is just the total probability of making an error. Therefore for this cost assignment the Bayes test is minimizing the total probability of error. The test is

$$\ln \Lambda(\mathbf{R}) \underset{H_0}{\overset{H_1}{\gtrless}} \ln \frac{P_0}{P_1} = \ln P_0 - \ln\,(1 - P_0). \tag{40}$$

When the two hypotheses are equally likely, the threshold is zero. This assumption is normally true in digital communication systems. These processors are commonly referred to as minimum probability of error receivers.

A second special case of interest arises when the a priori probabilities are unknown. To investigate this case we look at (8) again. We observe that once the decision regions Z_0 and Z_1 are chosen, the values of the ntegrals are determined. We denote these values in the following manner:

$$\begin{aligned} P_F &= \int_{Z_1} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R},\\ P_D &= \int_{Z_1} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R},\\ P_M &= \int_{Z_0} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R} = 1 - P_D. \end{aligned} \tag{41}$$

We see that these quantities are *conditional probabilities.* The subscripts are mnemonic and chosen from the radar problem in which hypothesis H_1 corresponds to the presence of a target and hypothesis H_0 corresponds to its absence. P_F is the probability of a *false alarm* (i.e., we say the target is present when it is not); P_D is the probability of *detection* (i.e., we say the target is present when it is); P_M is the probability of a *miss* (we say the target is absent when it is present). Although we are interested in a much larger class of problems than this notation implies, we shall use it for convenience.

For any choice of decision regions the risk expression in (8) can be written in the notation of (41):

$$\begin{aligned} \mathcal{R} = P_0 C_{10} + P_1 C_{11} + P_1(C_{01} - C_{11})P_M&\\ - P_0(C_{10} - C_{00})(1 - P_F).& \end{aligned} \tag{42}$$

Because

$$P_0 = 1 - P_1, \tag{43}$$

(42) becomes

$$\begin{aligned} \mathcal{R}(P_1) = C_{00}(1 - P_F) + C_{10}P_F&\\ + P_1[(C_{11} - C_{00}) + (C_{01} - C_{11})P_M - (C_{10} - C_{00})P_F].& \end{aligned} \tag{44}$$

Now, if all the costs and a priori probabilities are known, we can find a Bayes test. In Fig. 2.7*a* we plot the Bayes risk, $\mathcal{R}_B(P_1)$, as a function of P_1. Observe that as P_1 changes the decision regions for the Bayes test change and therefore P_F and P_M change.

Now consider the situation in which a certain P_1 (say $P_1 = P_1^*$) is *assumed* and the corresponding Bayes test designed. We now fix the threshold and assume that P_1 is allowed to change. We denote the risk for this fixed threshold test as $\mathcal{R}_F(P_1^*, P_1)$. Because the threshold is fixed, P_F and P_M are fixed, and (44) is just a straight line. Because it is a Bayes test for $P_1 = P_1^*$, it touches the $\mathcal{R}_B(P_1)$ curve at that point. Looking at (14), we see that the threshold changes continuously with P_1. Therefore, whenever $P_1 \neq P_1^*$, the threshold in the Bayes test will be different. Because the Bayes test minimizes the risk,

$$\mathcal{R}_F(P_1^*, P_1) \geq \mathcal{R}_B(P_1).$$

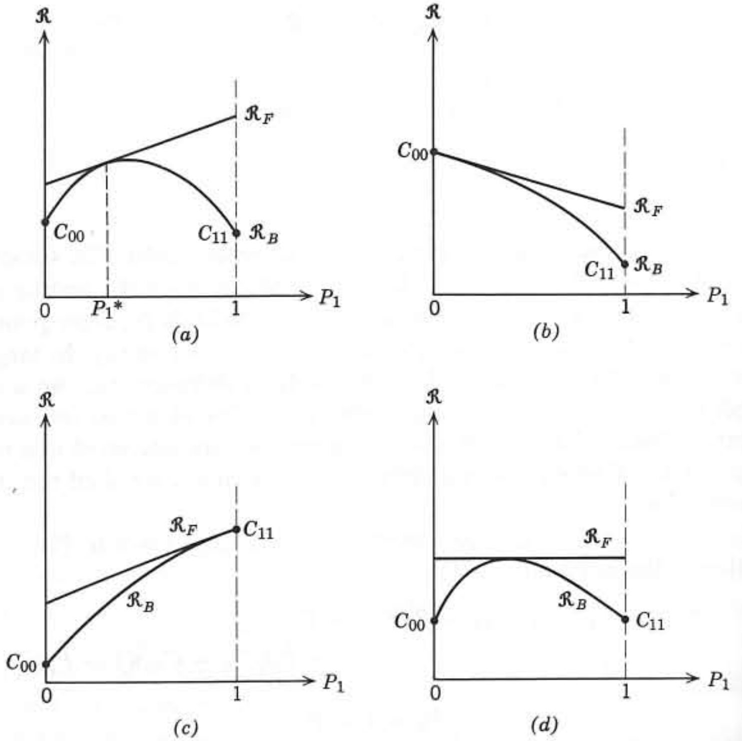


Fig. 2.7 Risk curves: (*a*) fixed risk versus typical Bayes risk; (*b*) maximum value of $\mathcal{R}_1$ at $P_1 = 0$.

If Λ is a continuous random variable with a probability distribution function that is strictly monotonic, then changing η always changes the risk. $\mathcal{R}_B(P_1)$ is strictly concave downward and the inequality in (45) is strict. This case, which is one of particular interest to us, is illustrated in Fig. 2.7*a*. We see that $\mathcal{R}_F(P_1^*, P_1)$ is tangent to $\mathcal{R}_B(P_1)$ at $P_1 = P_1^*$. These curves demonstrate the effect of incorrect knowledge of the a priori probabilities.

An interesting problem is encountered if we assume that the a priori probabilities are chosen to make our performance as bad as possible. In other words, P_1 is chosen to maximize our risk $\mathcal{R}_F(P_1^*, P_1)$. Three possible examples are given in Figs. 2.7*b*, *c*, and *d*. In Fig. 2.7*b* the maximum of $\mathcal{R}_B(P_1)$ occurs at $P_1 = 0$. To minimize the maximum risk we use a Bayes test designed assuming $P_1 = 0$. In Fig. 2.7*c* the maximum of $\mathcal{R}_B(P_1)$ occurs at $P_1 = 1$. To minimize the maximum risk we use a Bayes test designed assuming $P_1 = 1$. In Fig. 2.7*d* the maximum occurs inside the interval [0, 1], and we choose $\mathcal{R}_F$ to be the horizontal line. This implies that the coefficient of P_1 in (44) must be zero:

$$(C_{11} - C_{00}) + (C_{01} - C_{11})P_M - (C_{10} - C_{00})P_F = 0. \tag{46}$$

A Bayes test designed to minimize the maximum possible risk is called a *minimax test*. Equation 46 is referred to as the minimax equation and is useful whenever the maximum of $\mathcal{R}_B(P_1)$ is interior to the interval.

A special cost assignment that is frequently logical is

$$C_{00} = C_{11} = 0 \tag{47}$$

(This guarantees the maximum is interior.)

Denoting,

$$\begin{aligned} C_{01} &= C_M, \\ C_{10} &= C_F. \end{aligned} \tag{48}$$

the risk is,

$$\begin{aligned} \mathcal{R}_F &= C_F P_F + P_1(C_M P_M - C_F P_F) \\ &= P_0 C_F P_F + P_1 C_M P_M \end{aligned} \tag{49}$$

and the minimax equation is

$$C_M P_M = C_F P_F. \tag{50}$$

Before continuing our discussion of likelihood ratio tests we shall discuss a second criterion and prove that it also leads to a likelihood ratio test.

Neyman–Pearson Tests. In many physical situations it is difficult to assign realistic costs or a priori probabilities. A simple procedure to bypass this difficulty is to work with the *conditional probabilities* P_F and P_D. In general, we should like to make P_F as small as possible and P_D as large as possible. For most problems of practical importance these are conflicting objectives. An obvious criterion is to constrain one of the probabilities and maximize (or minimize) the other. A specific statement of this criterion is the following:

Neyman–Pearson Criterion. Constrain $P_F = \alpha' \le \alpha$ and design a test to maximize P_D (or minimize P_M) under this constraint.

The solution is obtained easily by using Lagrange multipliers. We construct the function F,

$$F = P_M + \lambda[P_F - \alpha'], \tag{51}$$

or

$$F = \int_{Z_0} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\,d\mathbf{R} + \lambda\left[\int_{Z_1} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\,d\mathbf{R} - \alpha'\right], \tag{52}$$

Clearly, if $P_F = \alpha'$, then minimizing F minimizes P_M.

or

$$F = \lambda(1 - \alpha') + \int_{Z_0} [p_{\mathbf{r}|H_1}(\mathbf{R}|H_1) - \lambda p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)]\, d\mathbf{R}. \tag{53}$$

Now observe that for any positive value of λ an LRT will minimize F. (A negative value of λ gives an LRT with the inequalities reversed.)

This follows directly, because to minimize F we assign a point $\mathbf{R}$ to Z_0 only when the term in the bracket is negative. This is equivalent to the test

$$\frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)} < \lambda, \qquad \text{assign point to } Z_0 \text{ or say } H_0. \tag{54}$$

The quantity on the left is just the likelihood ratio. Thus F is minimized by the likelihood ratio test

$$\Lambda(\mathbf{R}) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda. \tag{55}$$

To satisfy the constraint we choose λ so that $P_F = \alpha'$. If we denote the density of Λ when H_0 is true as $p_{\Lambda|H_0}(\Lambda|H_0)$, then we require

$$P_F = \int_\lambda^\infty p_{\Lambda|H_0}(\Lambda|H_0)\, d\Lambda = \alpha'. \tag{56}$$

Solving (56) for λ gives the threshold. The value of λ given by (56) will be non-negative because $p_{\Lambda|H_0}(\Lambda|H_0)$ is zero for negative values of λ. Observe that decreasing λ is equivalent to increasing Z_1, the region where we say H_1. Thus P_D increases as λ decreases. Therefore we decrease λ until we obtain the largest possible $\alpha' \le \alpha$. In most cases of interest to us P_F is a continuous function of λ and we have $P_F = \alpha$. We shall assume this continuity in all subsequent discussions. Under this assumption the Neyman–Pearson criterion leads to a likelihood ratio test. On p. 41 we shall see the effect of the continuity assumption not being valid.

Summary. In this section we have developed two ideas of fundamental importance in hypothesis testing. The first result is the demonstration that for a Bayes or a Neyman–Pearson criterion the optimum test consists of processing the observation $\mathbf{R}$ to find the likelihood ratio $\Lambda(\mathbf{R})$ and then comparing $\Lambda(\mathbf{R})$ to a threshold in order to make a decision. Thus, regardless of the dimensionality of the observation space, the decision space is one-dimensional.

The second idea is that of a sufficient statistic $l(\mathbf{R})$. The idea of a sufficient statistic originated when we constructed the likelihood ratio and saw that it depended explicitly only on $l(\mathbf{R})$. If we actually construct $\Lambda(\mathbf{R})$ and then recognize $l(\mathbf{R})$, the notion of a sufficient statistic is perhaps of secondary value. A more important case is when we can recognize $l(\mathbf{R})$ directly. An easy way to do this is to examine the geometric interpretation of a sufficient statistic. We considered the observations $r_1, r_2, \ldots, r_N$ as a point $\mathbf{r}$ in an N-dimensional space, and one way to describe this point is to use these coordinates. When we choose a sufficient statistic, we are simply describing the point in a coordinate system that is more useful for the decision problem. We denote the first coordinate in this system by l, the sufficient statistic, and the remaining $N - 1$ coordinates which will not affect our decision by the $(N - 1)$-dimensional vector $\mathbf{y}$. Thus

$$\Lambda(\mathbf{R}) = \Lambda(L, \mathbf{Y}) = \frac{p_{l,\mathbf{y}|H_1}(L, \mathbf{Y}|H_1)}{p_{l,\mathbf{y}|H_0}(L, \mathbf{Y}|H_0)}. \tag{57}$$

Now the expression on the right can be written as

$$\Lambda(L, \mathbf{Y}) = \frac{p_{l|H_1}(L|H_1) p_{\mathbf{y}|l,H_1}(\mathbf{Y}|L, H_1)}{p_{l|H_0}(L|H_0) p_{\mathbf{y}|l,H_0}(\mathbf{Y}|L, H_0)}. \tag{58}$$

If l is a sufficient statistic, then $\Lambda(\mathbf{R})$ must reduce to $\Lambda(L)$. This implies that the second terms in the numerator and denominator must be equal. In other words,

$$p_{\mathbf{y}|l,H_0}(\mathbf{Y}|L, H_0) = p_{\mathbf{y}|l,H_1}(\mathbf{Y}|L, H_1) \tag{59}$$

because the density of $\mathbf{y}$ cannot depend on which hypothesis is true. We see that choosing a sufficient statistic simply amounts to picking a coordinate system in which one coordinate contains all the information necessary to making a decision. The other coordinates contain no information and can be disregarded for the purpose of making a decision.

In Example 1 the new coordinate system could be obtained by a simple rotation. For example, when $N = 2$,

$$\begin{aligned} L &= \frac{1}{\sqrt{2}}(R_1 + R_2), \\ Y &= \frac{1}{\sqrt{2}}(R_1 - R_2). \end{aligned} \tag{60}$$

In Example 2 the new coordinate system corresponded to changing to polar coordinates. For $N = 2$

$$\begin{aligned} L &= R_1^2 + R_2^2, \\ Y &= \tan^{-1}\frac{R_2}{R_1}. \end{aligned} \tag{61}$$

Notice that the vector $\mathbf{y}$ can be chosen in order to make the demonstration of the condition in (59) as simple as possible. The only requirement is that the pair $(l, \mathbf{y})$ must describe any point in the observation space. We should also observe that the condition

$$p_{\mathbf{y}|H_1}(\mathbf{Y}|H_1) = p_{\mathbf{y}|H_0}(\mathbf{Y}|H_0) \tag{62}$$

does *not* imply (59) unless l and $\mathbf{y}$ are independent under H_1 and H_0. Frequently we will choose $\mathbf{y}$ to obtain this independence and then use (62) to verify that l is a sufficient statistic.

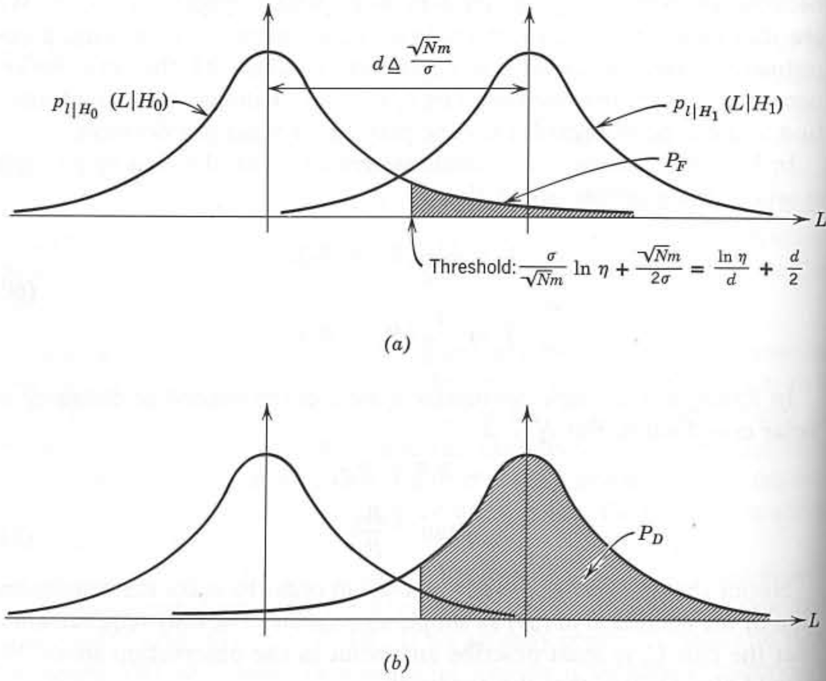
2.2.2 Performance: Receiver Operating Characteristic

To complete our discussion of the simple binary problem we must evaluate the performance of the likelihood ratio test. For a Neyman–Pearson test the values of P_F and P_D completely specify the test performance. Looking at (42) we see that the Bayes risk $\mathcal{R}_B$ follows easily if P_F and P_D are known. Thus we can concentrate our efforts on calculating P_F and P_D.

We begin by considering Example 1 in Section 2.2.1.

Example 1. From (25) we see that an equivalent test is

$$l = \frac{1}{\sqrt{N}\,\sigma} \sum_{i=1}^{N} R_i \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\sigma}{\sqrt{N}\,m} \ln \eta + \frac{\sqrt{N}\,m}{2\sigma}. \tag{63}$$


$d \triangleq \frac{\sqrt{N}m}{\sigma}$
$p_{l|H_0}\,(L|H_0)$
$p_{l|H_1}\,(L|H_1)$
P_F
L
Threshold: $\frac{\sigma}{\sqrt{N}m} \ln \eta + \frac{\sqrt{N}m}{2\sigma} = \frac{\ln \eta}{d} + \frac{d}{2}$
P_D
L

(a)

(b)

Fig. 2.8 Error probabilities: *(a)* P_F calculation; *(b)* P_D calculation.

We have multiplied (25) by $\sigma/\sqrt{N}\,m$ to normalize the next calculation. Under H_0, l is obtained by adding N independent zero-mean Gaussian variables with variance σ^2 and then dividing by $\sqrt{N}\,\sigma$. Therefore l is $N(0, 1)$.

Under H_1, l is $N(\sqrt{N}\,m/\sigma, 1)$. The probability densities on the two hypotheses are sketched in Fig. 2.8*a*. The threshold is also shown. Now, P_F is simply the integral of $p_{l|H_0}(L|H_0)$ to the right of the threshold.

Thus

$$P_F = \int_{(\ln \eta)/d + d/2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx, \tag{64}$$

where $d \triangleq \sqrt{N}\,m/\sigma$ is the distance between the means of the two densities. The integral in (64) is tabulated in many references (e.g., [3] or [4]).

We generally denote

$$\operatorname{erf}_*(X) \triangleq \int_{-\infty}^{X} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx, \tag{65}$$

where erf_* is an abbreviation for the error function† and

$$\operatorname{erfc}_*(X) \triangleq \int_{X}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \tag{66}$$

is its complement. In this notation

$$P_F = \operatorname{erfc}_*\left(\frac{\ln \eta}{d} + \frac{d}{2}\right). \tag{67}$$

Similarly, P_D is the integral of $p_{l|H_1}(L|H_1)$ to the right of the threshold, as shown in Fig. 2.8*b*:

$$\begin{aligned} P_D &= \int_{(\ln \eta)/d + d/2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-d)^2}{2}\right] dx \\ &= \int_{(\ln \eta)/d - d/2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy \triangleq \operatorname{erfc}_*\left(\frac{\ln \eta}{d} - \frac{d}{2}\right). \end{aligned} \tag{68}$$

In Fig. 2.9*a* we have plotted P_D versus P_F for various values of d with η as the varying parameter. For $\eta = 0$, $\ln \eta = -\infty$, and the processor always guesses H_1. Thus $P_F = 1$ and $P_D = 1$. As η increases, P_F and P_D decrease. When $\eta = \infty$, the processor always guesses H_0 and $P_F = P_D = 0$.

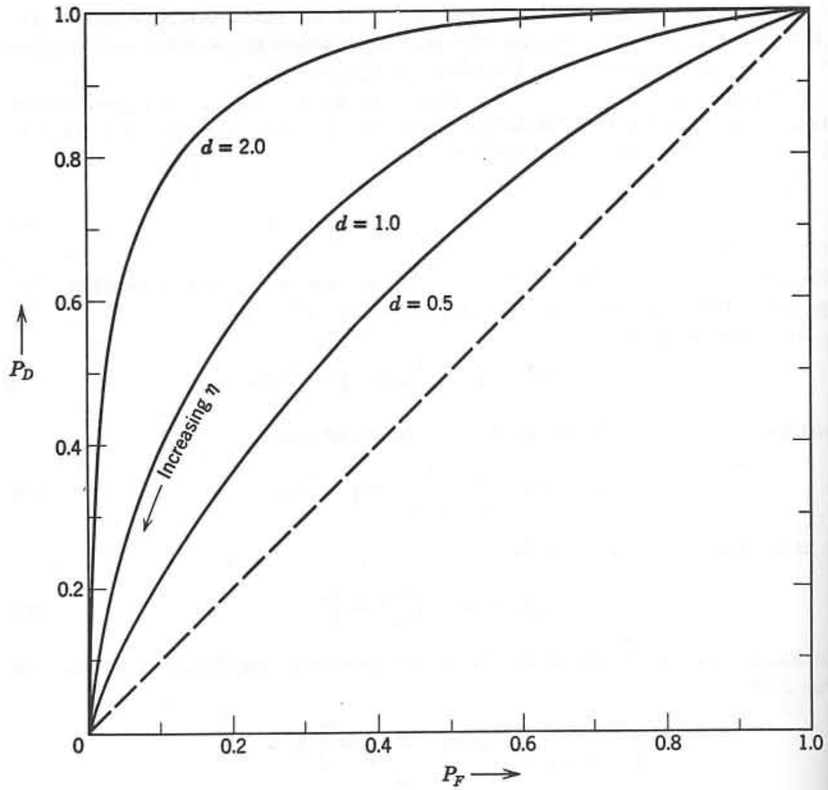
As we would expect from Fig. 2.8, the performance increases monotonically with d. In Fig. 2.9*b* we have replotted the results to give P_D versus d with P_F as a parameter on the curves. For a particular d we can obtain any point on the curve by choosing η appropriately $(0 \le \eta \le \infty)$.

The result in Fig. 2.9*a* is referred to as the receiver operating characteristic (ROC). It completely describes the performance of the test as a function of the parameter of interest.

A special case that will be important when we look at communication systems is the case in which we want to minimize the total probability of error

$$\Pr(\epsilon) \triangleq P_0 P_F + P_1 P_M. \tag{69a}$$

† The function that is usually tabulated is $\operatorname{erf}(X) = \sqrt{2/\pi} \int_0^X \exp(-y^2)\, dy$, which is related to (65) in an obvious way.



(a)

Fig. 2.9 (*a*) Receiver operating characteristic: Gaussian variables with unequal means.

The threshold for this criterion was given in (40). For the special case in which $P_0 = P_1$ the threshold η equals one and

$$\Pr(\epsilon) = \tfrac{1}{2}(P_F + P_M). \tag{69b}$$

Using (67) and (68) in (69), we have

$$\Pr(\epsilon) = \int_{+d/2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = \operatorname{erfc}_*\left(+\frac{d}{2}\right). \tag{70}$$

It is obvious from (70) that we could also obtain the Pr (ϵ) from the ROC. However, if this is the only threshold setting of interest, it is generally easier to calculate the Pr (ϵ) directly.

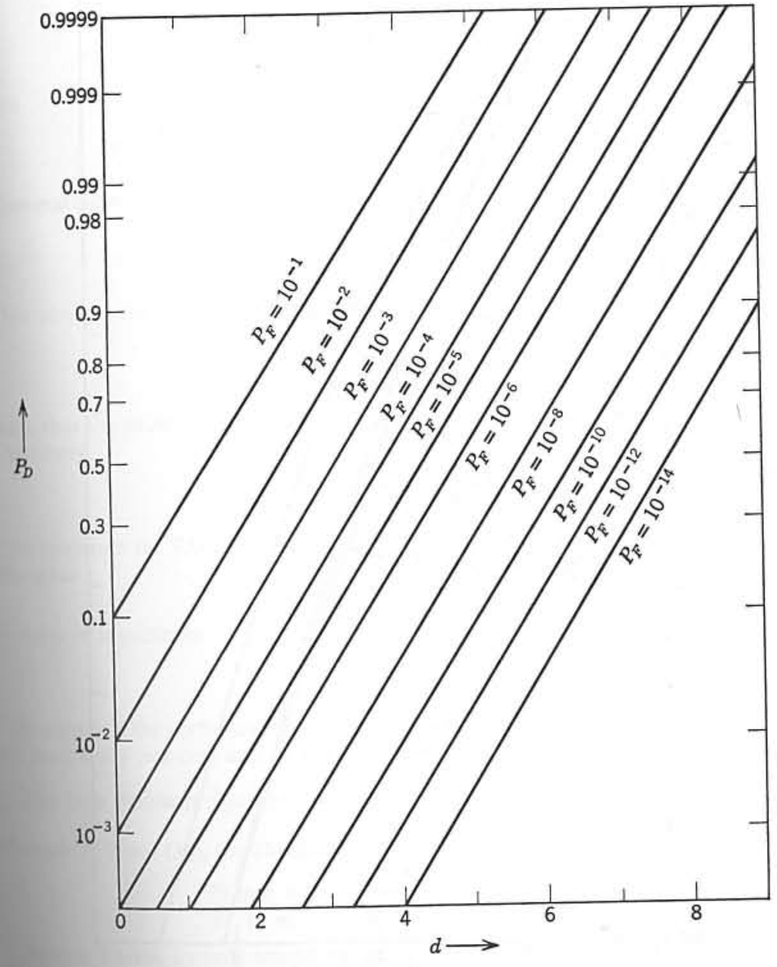
Before calculating the performance of the other two examples, it is worthwhile to point out two simple bounds on $\operatorname{erfc}_*(X)$. They will enable

us to discuss its approximate behavior analytically. For $X > 0$

$$\frac{1}{\sqrt{2\pi}\,X}\left(1 - \frac{1}{X^2}\right) \exp\left(-\frac{X^2}{2}\right) < \operatorname{erfc}_*(X) < \frac{1}{\sqrt{2\pi}\,X} \exp\left(-\frac{X^2}{2}\right). \tag{71}$$

This can be derived by integrating by parts. (See Problem 2.2.15 or Feller [30].) A second bound is

$$\operatorname{erfc}_*(X) < \tfrac{1}{2} \exp\left(-\frac{X^2}{2}\right), \qquad x > 0, \tag{72}$$



(b)

Fig. 2.9 (*b*) detection probability versus *d*.

which can also be derived easily (see Problem 2.2.16). The four curves are plotted in Fig. 2.10. We note that $\text{erfc}_*(X)$ decreases exponentially.

The receiver operating characteristics for the other two examples are also of interest.

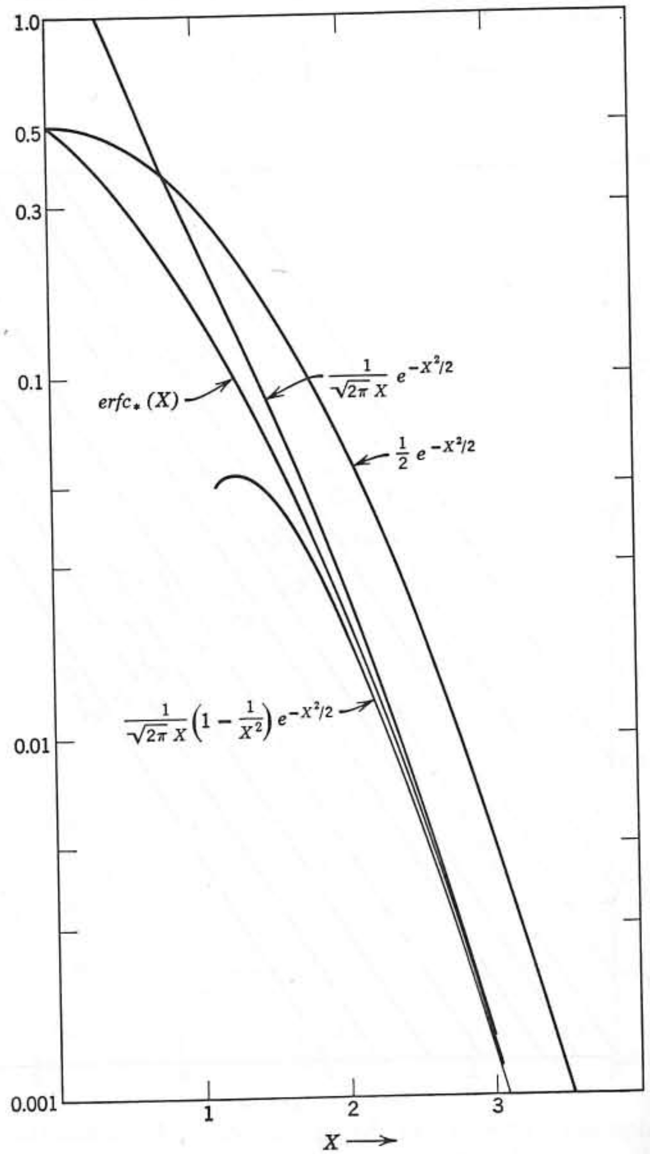


Fig. 2.10 Plot of $\text{erfc}_*(X)$ and related functions.

Example 2. In this case the test is

$$l(\mathbf{R}) = \sum_{i=1}^{N} R_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}\left(\ln \eta - N \ln \frac{\sigma_0}{\sigma_1}\right) = \gamma, \qquad (\sigma_1 > \sigma_0). \tag{73}$$

The performance calculation for arbitrary N is somewhat tedious, so we defer it until Section 2.6. A particularly simple case appearing frequently in practice is $N = 2$. Under H_0 the r_i are independent zero-mean Gaussian variables with variances equal to σ_0^2:

$$P_F = \Pr(l \geq \gamma | H_0) = \Pr(r_1^2 + r_2^2 \geq \gamma | H_0). \tag{74}$$

To evaluate the expression on the right, we change to polar coordinates:

$$\begin{aligned} r_1 &= z \cos \theta, \qquad & z &= \sqrt{r_1^2 + r_2^2} \\ r_2 &= z \sin \theta, \qquad & \theta &= \tan^{-1} \frac{r_2}{r_1} \end{aligned} \tag{75}$$

Then

$$\Pr(z^2 \geq \gamma | H_0) = \int_0^{2\pi} d\theta \int_{\sqrt{\gamma}}^{\infty} Z \frac{1}{2\pi\sigma_0^2} \exp\left(-\frac{Z^2}{2\sigma_0^2}\right) dZ. \tag{76}$$

Integrating with respect to θ, we have

$$P_F = \int_{\sqrt{\gamma}}^{\infty} Z \frac{1}{\sigma_0^2} \exp\left(-\frac{Z^2}{2\sigma_0^2}\right) dZ. \tag{77}$$

We observe that l, the sufficient statistic, equals z^2. Changing variables, we have

$$P_F = \int_{\gamma}^{\infty} \frac{1}{2\sigma_0^2} \exp\left(-\frac{L}{2\sigma_0^2}\right) dL = \exp\left(-\frac{\gamma}{2\sigma_0^2}\right). \tag{78}$$

(Note that the probability density of the sufficient statistic is exponential.)

Similarly,

$$P_D = \exp\left(-\frac{\gamma}{2\sigma_1^2}\right). \tag{79}$$

To construct the ROC we can combine (78) and (79) to eliminate the threshold γ. This gives

$$P_D = (P_F)^{\sigma_0^2/\sigma_1^2}. \tag{80}$$

In terms of logarithms

$$\ln P_D = \frac{\sigma_0^2}{\sigma_1^2} \ln P_F. \tag{81}$$

As expected, the performance improves monotonically as the ratio σ_1^2/σ_0^2 increases. We shall study this case and its generalizations in more detail in Section 2.6.

The two Poisson distributions are the third example.

Example 3. From (38), the likelihood ratio test is

$$n \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\ln \eta + m_1 - m_0}{\ln m_1 - \ln m_0} = \gamma, \qquad (m_1 > m_0). \tag{82}$$

Because n takes on only integer values, it is more convenient to rewrite (82) as

$$n \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_I, \qquad \gamma_I = 0, 1, 2, \ldots,$$

where γ_I takes on only integer values. Using (35),

$$P_D = 1 - e^{-m_1} \sum_{n=0}^{\gamma_I - 1} \frac{(m_1)^n}{n!}, \qquad \gamma_I = 0, 1, 2, \ldots, \tag{84}$$

and from (36)

$$P_F = 1 - e^{-m_0} \sum_{n=0}^{\gamma_I - 1} \frac{(m_0)^n}{n!}, \qquad \gamma_I = 0, 1, 2, \ldots. \tag{85}$$

The resulting ROC is plotted in Fig. 2.11*a* for some representative values of m_0 and m_1.

We see that it consists of a series of points and that P_F goes from 1 to $1 - e^{-m_0}$ when the threshold is changed from 0 to 1. Now suppose we wanted P_F to have an intermediate value, say $1 - \frac{1}{2}e^{-m_0}$. To achieve this performance we proceed in the following manner. Denoting the LRT with $\gamma_I = 0$ as LRT No. 0 and the LRT with $\gamma_I = 1$ as LRT No. 1, we have the following table:

LRT	γ_I	P_F	P_D
0	0	1	1
1	1	$1 - e^{-m_0}$	$1 - e^{-m_1}$

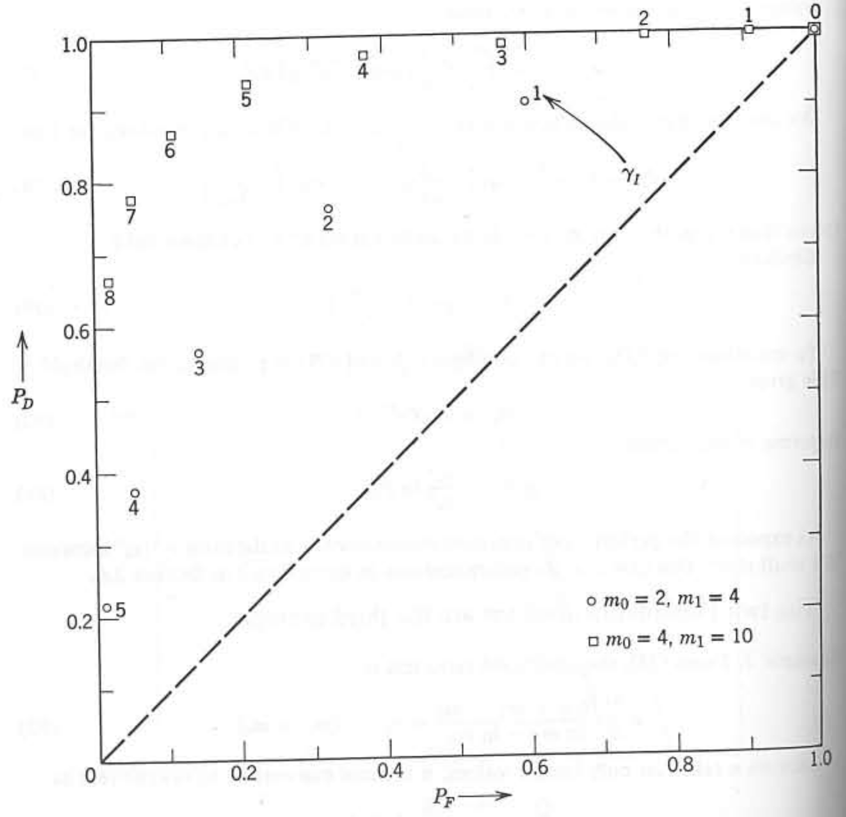


Fig. 2.11 (*a*) Receiver operating characteristic, Poisson problem.

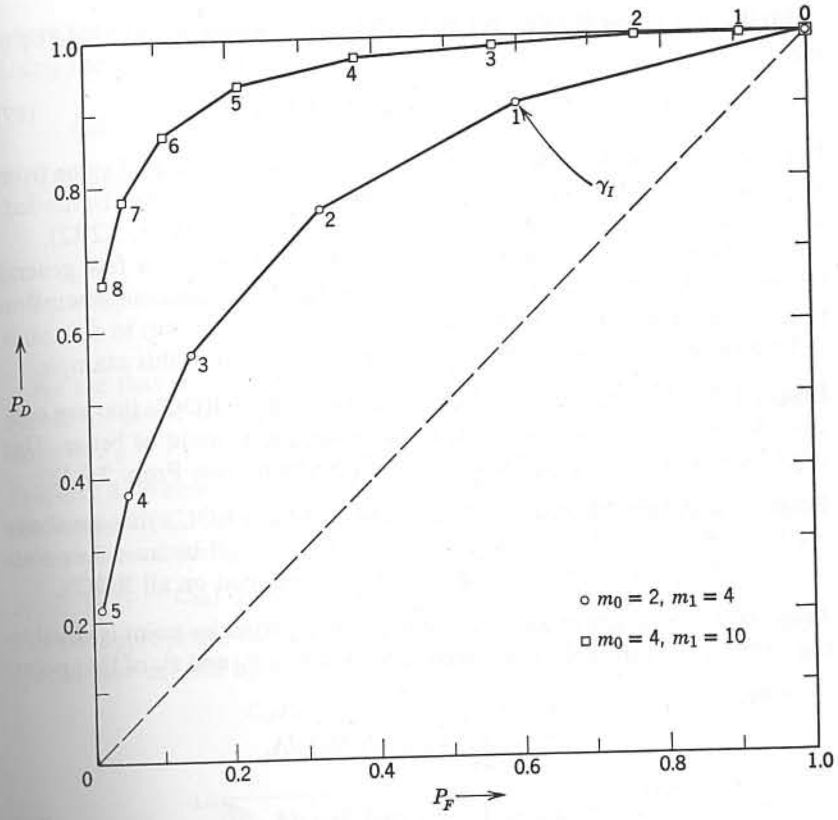


Fig. 2.11 (*b*) Receiver operating characteristic with randomized decision rule.

To get the desired value of P_F we use LRT No. 0 with probability $\frac{1}{2}$ and LRT No. 1 with probability $\frac{1}{2}$. The test is

$$\begin{aligned} \text{If } n = 0, \quad & \text{say } H_1 \text{ with probability } \tfrac{1}{2}, \\ & \text{say } H_0 \text{ with probability } \tfrac{1}{2}, \\ n \geq 1 \quad & \text{say } H_1. \end{aligned}$$

This procedure, in which we mix two likelihood ratio tests in some probabilistic manner, is called a *randomized decision rule*. The resulting P_D is simply a weighted combination of detection probabilities for the two tests.

$$P_D = 0.5(1) + 0.5(1 - e^{-m_1}) = (1 - 0.5\, e^{-m_1}). \tag{86}$$

We see that the ROC for randomized tests consists of straight lines which connect the points in Fig. 2.11*a*, as shown in Fig. 2.11*b*. The reason that we encounter a randomized test is that the observed random variables are discrete. Therefore $\Lambda(\mathbf{R})$ is a discrete random variable and, using an ordinary likelihood ratio test, only certain values of P_F are possible.

Looking at the expression for P_F in (56) and denoting the threshold by η, we have

$$P_F(\eta) = \int_{\eta}^{\infty} p_{\Lambda|H_0}(X|H_0)\, dX. \tag{87}$$

If $P_F(\eta)$ is a continuous function of η, we can achieve a desired value from 0 to 1 by a suitable choice of η and a randomized test will never be needed. This is the only case of interest to us in the sequel (see Prob. 2.2.12).

With these examples as a background, we now derive a few general properties of receiver operating characteristics. We confine our discussion to continuous likelihood ratio tests.

Two properties of *all* ROC's follow immediately from this example.

Property 1. All continuous likelihood ratio tests have ROC's that are concave downward. If they were not, a randomized test would be better. This would contradict our proof that a LRT is optimum (see Prob. 2.2.12).

Property 2. All continuous likelihood ratio tests have ROC's that are above the $P_D = P_F$ line. This is just a special case of Property 1 because the points ($P_F = 0, P_D = 0$) and ($P_F = 1, P_D = 1$) are contained on all ROC's.

Property 3. The slope of a curve in a ROC at a particular point is equal to the value of the threshold η required to achieve the P_D and P_F of that point.

Proof.

$$P_D = \int_{\eta}^{\infty} p_{\Lambda|H_1}(\Lambda|H_1)\, d\Lambda,$$

$$P_F = \int_{\eta}^{\infty} p_{\Lambda|H_0}(\Lambda|H_0)\, d\Lambda. \tag{88}$$

Differentiating both expressions with respect to η and writing the results as a quotient, we have

$$\frac{dP_D/d\eta}{dP_F/d\eta} = \frac{-p_{\Lambda|H_1}(\eta|H_1)}{-p_{\Lambda|H_0}(\eta|H_0)} = \frac{dP_D}{dP_F}. \tag{89}$$

We now show that

$$\frac{p_{\Lambda|H_1}(\eta|H_1)}{p_{\Lambda|H_0}(\eta|H_0)} = \eta. \tag{90}$$

Let

$$\Omega(\eta) \triangleq \{\mathbf{R}|\Lambda(\mathbf{R}) \geq \eta\} = \left[\mathbf{R}\left|\frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)} \geq \eta\right.\right]. \tag{91}$$

Then

$$\begin{aligned} P_D(\eta) \triangleq \Pr\{\Lambda(\mathbf{R}) \geq \eta|H_1\} &= \int_{\Omega(\eta)} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R} \\ &= \int_{\Omega(\eta)} \Lambda(\mathbf{R}) p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R}, \end{aligned} \tag{92}$$

where the last equality follows from the definition of the likelihood ratio. Using the definition of $\Omega(\eta)$, we can rewrite the last integral

$$P_D(\eta) = \int_{\Omega(\eta)} \Lambda(\mathbf{R}) p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} = \int_{\eta}^{\infty} X p_{\Lambda|H_0}(X|H_0)\, dX. \tag{93}$$

Differentiating (93) with respect to η, we obtain

$$\frac{dP_D(\eta)}{d\eta} = -\eta p_{\Lambda|H_0}(\eta|H_0). \tag{94}$$

Equating the expression for $dP_D(n)/d\eta$ in the numerator of (89) to the right side of (94) gives the desired result.

We see that this result is consistent with Example 1. In Fig. 2.9*a*, the curves for nonzero d have zero slope at $P_F = P_D = 1$ ($\eta = 0$) and infinite slope at $P_F = P_D = 0$ ($\eta = \infty$).

Property 4. Whenever the maximum value of the Bayes risk is interior to the interval (0, 1) on the P_1 axis, the minimax operating point is the intersection of the line

$$(C_{11} - C_{00}) + (C_{01} - C_{11})(1 - P_D) - (C_{10} - C_{00})P_F = 0 \tag{95}$$

and the appropriate curve of the ROC (see 46). In Fig. 2.12 we show the special case defined by (50),

$$C_F P_F = C_M P_M = C_M(1 - P_D), \tag{96}$$

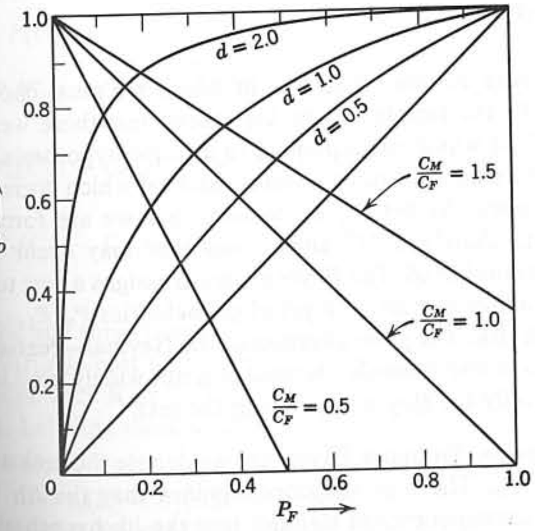


Fig. 2.12 Determination of minimax operating point.

superimposed on the ROC of Example 1. We see that it starts at the point $P_F = 0$, $P_D = 1$, and intersects the $P_F = 1$ line at

$$P_F = 1 - \frac{C_F}{C_M}. \tag{97}$$

This completes our discussion of the binary hypothesis testing problem. Several key ideas should be re-emphasized:

1. Using either a Bayes criterion or a Neyman–Pearson criterion, we find that the optimum test is a likelihood ratio test. Thus, regardless of the dimensionality of the observation space, the test consists of comparing a scalar variable $\Lambda(\mathbf{R})$ with a threshold. (We assume $P_F(\eta)$ is continuous.)

2. In many cases construction of the LRT can be simplified if we can identify a sufficient statistic. Geometrically, this statistic is just that coordinate in a suitable coordinate system which describes the observation space that contains *all* the information necessary to make a decision.

3. A complete description of the LRT performance was obtained by plotting the conditional probabilities P_D and P_F as the threshold η was varied. The resulting ROC could be used to calculate the Bayes risk for any set of costs. In many cases only one value of the threshold is of interest and a complete ROC is not necessary.

A number of interesting binary tests are developed in the problems.

2.3 M HYPOTHESES

The next case of interest is one in which we must choose one of M hypotheses. In the simple binary hypothesis test there were two source outputs, each of which corresponded to a single hypothesis. In the simple M-ary test there are M source outputs, each of which corresponds to one of M hypotheses. As before, we assume that we are forced to make a decision. Thus there are M^2 alternatives that may occur each time the experiment is conducted. The Bayes criterion assigns a cost to each of these alternatives, assumes a set of a priori probabilities $P_0, P_1, \ldots, P_{M-1}$, and minimizes the risk. The generalization of the Neyman–Pearson criterion to M hypotheses is also possible. Because it is not widely used in practice, we shall discuss only the Bayes criterion in the text.

Bayes Criterion. To find a Bayes test we denote the cost of each course of action as C_{ij}. The first subscript signifies that the ith hypothesis is chosen. The second subscript signifies that the jth hypothesis is true. We denote the region of the observation space in which we choose H_i as Z_i and the a priori probabilities are P_i. The model is shown in Fig. 2.13. The expression for the risk is

$$\mathcal{R} = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} P_j C_{ij} \int_{Z_i} p_{\mathbf{r}|H_j}(\mathbf{R}|H_j)\, d\mathbf{R}. \tag{98}$$

To find the optimum Bayes test we simply vary the Z_i to minimize $\mathcal{R}$. This is a straightforward extension of the technique used in the binary case. For simplicity of notation, we shall only consider the case in which $M = 3$ in the text.

Noting that $Z_0 = Z - Z_1 - Z_2$, because the regions are disjoint, we obtain

$$\begin{aligned}\mathcal{R} = {} & P_0C_{00}\int_{Z-Z_1-Z_2} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\,d\mathbf{R} + P_0C_{10}\int_{Z_1} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\,d\mathbf{R}\\ & + P_0C_{20}\int_{Z_2} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\,d\mathbf{R} + P_1C_{11}\int_{Z-Z_0-Z_2} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\,d\mathbf{R}\\ & + P_1C_{01}\int_{Z_0} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\,d\mathbf{R} + P_1C_{21}\int_{Z_2} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\,d\mathbf{R}\\ & + P_2C_{22}\int_{Z-Z_0-Z_1} p_{\mathbf{r}|H_2}(\mathbf{R}|H_2)\,d\mathbf{R} + P_2C_{02}\int_{Z_0} p_{\mathbf{r}|H_2}(\mathbf{R}|H_2)\,d\mathbf{R}\\ & + P_2C_{12}\int_{Z_1} p_{\mathbf{r}|H_2}(\mathbf{R}|H_2)\,d\mathbf{R}.\end{aligned} \tag{99}$$

This reduces to

$$\begin{aligned}\mathcal{R} = {} & P_0C_{00} + P_1C_{11} + P_2C_{22}\\ & + \int_{Z_0} [P_2(C_{02} - C_{22})p_{\mathbf{r}|H_2}(\mathbf{R}|H_2) + P_1(C_{01} - C_{11})p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)]\,d\mathbf{R}\\ & + \int_{Z_1} [P_0(C_{10} - C_{00})p_{\mathbf{r}|H_0}(\mathbf{R}|H_0) + P_2(C_{12} - C_{22})p_{\mathbf{r}|H_2}(\mathbf{R}|H_2)]\,d\mathbf{R}\\ & + \int_{Z_2} [P_0(C_{20} - C_{00})p_{\mathbf{r}|H_0}(\mathbf{R}|H_0) + P_1(C_{21} - C_{11})p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)]\,d\mathbf{R}.\end{aligned} \tag{100}$$

As before, the first three terms represent the fixed cost and the integrals represent the variable cost that depends on our choice of Z_0, Z_1, and Z_2. Clearly, we assign each $\mathbf{R}$ to the region in which the value of the integrand is the smallest. Labeling these integrands $I_0(\mathbf{R})$, $I_1(\mathbf{R})$, and $I_2(\mathbf{R})$, we have the following rule:

$$\begin{aligned}&\text{if } I_0(\mathbf{R}) < I_1(\mathbf{R}) \text{ and } I_2(\mathbf{R}), \text{ choose } H_0,\\ &\text{if } I_1(\mathbf{R}) < I_0(\mathbf{R}) \text{ and } I_2(\mathbf{R}), \text{ choose } H_1,\\ &\text{if } I_2(\mathbf{R}) < I_0(\mathbf{R}) \text{ and } I_1(\mathbf{R}), \text{ choose } H_2.\end{aligned} \tag{101}$$

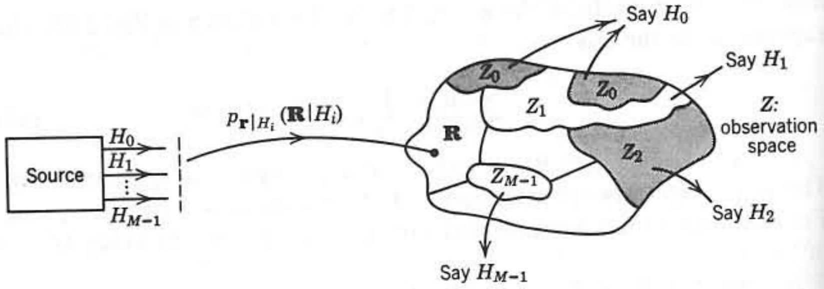


Fig. 2.13 ***M* hypothesis problem.**

We can write these terms in terms of likelihood ratios by defining

$$\Lambda_1(\mathbf{R}) \triangleq \frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)},$$
$$\Lambda_2(\mathbf{R}) \triangleq \frac{p_{\mathbf{r}|H_2}(\mathbf{R}|H_2)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)}. \tag{102}$$

Using (102) in (100) and (101), we have

$$P_1(C_{01} - C_{11})\,\Lambda_1(\mathbf{R}) \underset{H_0 \text{ or } H_2}{\overset{H_1 \text{ or } H_2}{\gtrless}} P_0(C_{10} - C_{00}) + P_2(C_{12} - C_{02})\,\Lambda_2(\mathbf{R}) \tag{103}$$

$$P_2(C_{02} - C_{22})\,\Lambda_2(\mathbf{R}) \underset{H_0 \text{ or } H_1}{\overset{H_2 \text{ or } H_1}{\gtrless}} P_0(C_{20} - C_{00}) + P_1(C_{21} - C_{01})\,\Lambda_1(\mathbf{R}), \tag{104}$$

$$P_2(C_{12} - C_{22})\,\Lambda_2(\mathbf{R}) \underset{H_1 \text{ or } H_0}{\overset{H_2 \text{ or } H_0}{\gtrless}} P_0(C_{20} - C_{10}) + P_1(C_{21} - C_{11})\,\Lambda_1(\mathbf{R}). \tag{105}$$

We see that the decision rules correspond to three lines in the Λ_1, Λ_2 plane. It is easy to verify that these lines intersect at a common point and therefore uniquely define three decision regions, as shown in Fig. 2.14. The decision space is two-dimensional for the three-hypothesis problem. It is easy to verify that M hypotheses *always* lead to a decision space which has, at most, $(M - 1)$ dimensions.

Several special cases will be useful in our later work. The first is defined by the assumptions

$$\begin{aligned} C_{00} = C_{11} &= C_{22} = 0, \\ C_{ij} &= 1, \qquad i \neq j. \end{aligned} \tag{106}$$

These equations indicate that any error is of equal importance. Looking at (98), we see that this corresponds to minimizing the total probability of error.

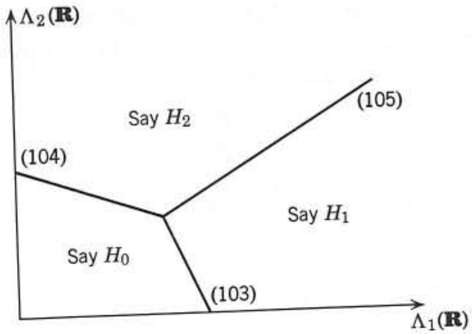
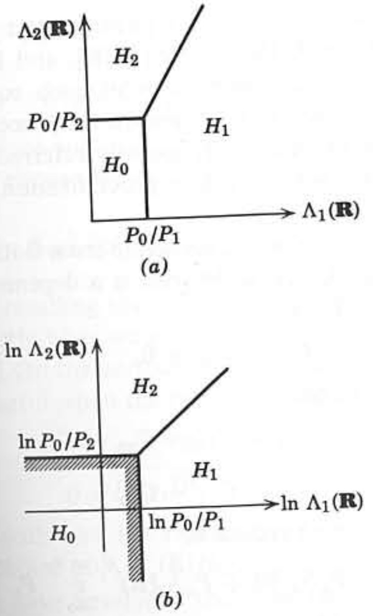


Fig. 2.14 Decision space.

Substituting into (103)–(105), we have

$$\begin{aligned} P_1\Lambda_1(\mathbf{R}) &\underset{H_0 \text{ or } H_2}{\overset{H_1 \text{ or } H_2}{\gtrless}} P_0, \\ P_2\Lambda_2(\mathbf{R}) &\underset{H_0 \text{ or } H_1}{\overset{H_2 \text{ or } H_1}{\gtrless}} P_0, \\ P_2\Lambda_2(\mathbf{R}) &\underset{H_1 \text{ or } H_0}{\overset{H_2 \text{ or } H_0}{\gtrless}} P_1\Lambda_1(\mathbf{R}). \end{aligned} \tag{107}$$



(*a*)

(*b*)

Fig. 2.15 Decision spaces.

The decision regions in the (Λ_1, Λ_2) plane are shown in Fig. 2.15*a*. In this particular case, the transition to the $(\ln \Lambda_1, \ln \Lambda_2)$ plane is straightforward (Fig. 2.15*b*). The equations are

$$\begin{aligned} \ln \Lambda_1(\mathbf{R}) &\underset{H_0 \text{ or } H_2}{\overset{H_1 \text{ or } H_2}{\gtrless}} \ln \frac{P_0}{P_1}, \\ \ln \Lambda_2(\mathbf{R}) &\underset{H_0 \text{ or } H_1}{\overset{H_1 \text{ or } H_2}{\gtrless}} \ln \frac{P_0}{P_2}, \\ \ln \Lambda_2(\mathbf{R}) &\underset{H_0 \text{ or } H_1}{\overset{H_0 \text{ or } H_2}{\gtrless}} \ln \Lambda_1(\mathbf{R}) + \ln \frac{P_1}{P_2}. \end{aligned} \tag{108}$$

The expressions in (107) and (108) are adequate, but they obscure an important interpretation of the processor. The desired interpretation is obtained by a little manipulation.

Substituting (102) into (103–105) and multiplying both sides by $p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)$, we have

$$\begin{aligned} P_1 p_{\mathbf{r}|H_1}(\mathbf{R}|H_1) &\underset{H_0 \text{ or } H_2}{\overset{H_1 \text{ or } H_2}{\gtrless}} P_0 p_{\mathbf{r}|H_0}(\mathbf{R}|H_0), \\ P_2 p_{\mathbf{r}|H_2}(\mathbf{R}|H_2) &\underset{H_0 \text{ or } H_1}{\overset{H_2 \text{ or } H_1}{\gtrless}} P_0 p_{\mathbf{r}|H_0}(\mathbf{R}|H_0), \\ P_2 p_{\mathbf{r}|H_2}(\mathbf{R}|H_2) &\underset{H_1 \text{ or } H_0}{\overset{H_2 \text{ or } H_0}{\gtrless}} P_1 p_{\mathbf{r}|H_1}(\mathbf{R}|H_1). \end{aligned} \tag{109}$$

Looking at (109), we see that an equivalent test is to compute the a posteriori probabilities Pr $[H_0|\mathbf{R}]$, Pr $[H_1|\mathbf{R}]$, and Pr $[H_2|\mathbf{R}]$ and choose the largest. (Simply divide both sides of each equation by $p_{\mathbf{r}}(\mathbf{R})$ and examine the resulting test.) For this reason the processor for the minimum probability of error criterion is frequently referred to as a *maximum a posteriori probability computer*. The generalization to M hypotheses is straightforward.

The next two topics deal with degenerate tests. Both results will be useful in later applications. A case of interest is a degenerate one in which we combine H_1 and H_2. Then

$$C_{12} = C_{21} = 0, \tag{110}$$

and, for simplicity, we can let

$$C_{01} = C_{10} = C_{20} = C_{02} \tag{111}$$

and

$$C_{00} = C_{11} = C_{22} = 0. \tag{112}$$

Then (103) and (104) both reduce to

$$P_1\Lambda_1(\mathbf{R}) + P_2\Lambda_2(\mathbf{R}) \underset{H_0}{\overset{H_1 \text{ or } H_2}{\gtrless}} P_0 \tag{113}$$

and (105) becomes an identity.

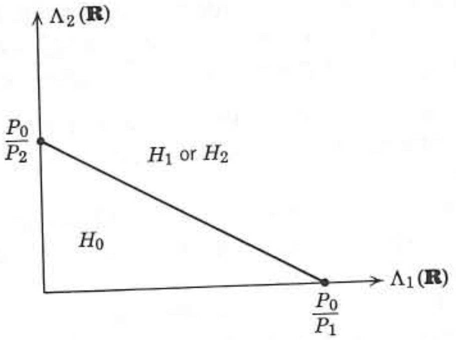


Fig. 2.16 Decision spaces.

The decision regions are shown in Fig. 2.16. Because we have eliminated all of the cost effect of a decision between H_1 and H_2, we have reduced it to a binary problem.

We next consider the dummy hypothesis technique. A simple example illustrates the idea. The actual problem has two hypotheses, H_1 and H_2, but occasionally we can simplify the calculations by introducing a dummy hypothesis H_0 which occurs with zero probability. We let

$$P_0 = 0, \qquad P_1 + P_2 = 1, \tag{114}$$

and

$$C_{12} = C_{02}, C_{21} = C_{01}.$$

Substituting these values into (103–105), we find that (103) and (104) imply that we always choose H_1 or H_2 and the test reduces to

$$P_2(C_{12} - C_{22})\,\Lambda_2(\mathbf{R}) \underset{H_1}{\overset{H_2}{\gtrless}} P_1(C_{21} - C_{11})\,\Lambda_1(\mathbf{R}). \tag{115}$$

Looking at (12) and recalling the definition of $\Lambda_1(\mathbf{R})$ and $\Lambda_2(\mathbf{R})$, we see that this result is exactly what we would expect. [Just divide both sides of (12) by $p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)$.] On the surface this technique seems absurd, but it will turn out to be useful when the ratio

$$\frac{p_{\mathbf{r}|H_2}(\mathbf{R}|H_2)}{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}$$

is difficult to work with and the ratios $\Lambda_1(\mathbf{R})$ and $\Lambda_2(\mathbf{R})$ can be made simple by a proper choice of $p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)$.

In this section we have developed the basic results needed for the M-hypothesis problem. We have not considered any specific exa because the details involved in constructing the likelihood ratios are the same as those in the binary case. Typical examples are given in the problems. Several important points should be emphasized.

1. The minimum dimension of the decision space is no more than $M - 1$. The boundaries of the decision regions are hyperplanes in the $(\Lambda_1, \ldots, \Lambda_{M-1})$ plane.
2. The optimum test is straightforward to find. We shall find however, when we consider specific examples that the error probabilities are frequently difficult to compute.
3. A particular test of importance is the minimum total probability of error test. Here we compute the a posteriori probability of each hypothesis $\Pr(H_i|\mathbf{R})$ and choose the largest.

These points will be appreciated more fully as we proceed through various applications.

These two sections complete our discussion of simple hypothesis tests. A case of importance that we have not yet discussed is the one in which several source outputs are combined to give a single hypothesis. To study this detection problem, we shall need some ideas from estimation theory. Therefore we defer the composite hypothesis testing problem until Section 2.5 and study the estimation problem next.

2.4 ESTIMATION THEORY

In the last two sections we have considered a problem in which one of several hypotheses occurred. As the result of a particular hypothesis, a vector random variable $\mathbf{r}$ was observed. Based on our observation, we shall try to choose the true hypothesis.

In this section we discuss the problem of *parameter estimation*. Before formulating the general problem, let us consider a simple example.

Example 1. We want to measure a voltage a at a single time instant. From physical considerations, we know that the voltage is between $-V$ and $+V$ volts. The measurement is corrupted by noise which may be modeled as an independent additive zero-mean Gaussian random variable n. The observed variable is r. Thus

$$r = a + n. \tag{116}$$

The probability density governing the observation process is $p_{r|a}(R|A)$. In this case

$$p_{r|a}(R|A) = p_n(R - A) = \frac{1}{\sqrt{2\pi}\,\sigma_n} \exp\left(-\frac{(R - A)^2}{2\sigma_n^2}\right). \tag{117}$$

The problem is to observe r and estimate a.

This example illustrates the basic features of the estimation problem.

A model of the general estimation problem is shown in Fig. 2.17. The model has the following four components:

Parameter Space. The output of the source is a parameter (or variable). We view this output as a point in a parameter space. For the single-parameter case, which we shall study first, this will correspond to segments of the line $-\infty < A < \infty$. In the example considered above the segment is $(-V, V)$.

Probabilistic Mapping from Parameter Space to Observation Space. This is the probability law that governs the effect of a on the observation.

Observation Space. In the classical problem this is a finite-dimensional space. We denote a point in it by the vector $\mathbf{R}$.

Estimation Rule. After observing $\mathbf{R}$, we shall want to estimate the value of a. We denote this estimate as $\hat{a}(\mathbf{R})$. This mapping of the observation space into an estimate is called the estimation rule. The purpose of this section is to investigate various estimation rules and their implementations.

The second and third components are familiar from the detection problem. The new features are the parameter space and the estimation rule. When we try to describe the parameter space, we find that two cases arise. In the first, the parameter is a random variable whose behavior is governed by a probability density. In the second, the parameter is an unknown quantity but not a random variable. These two cases are analogous to the

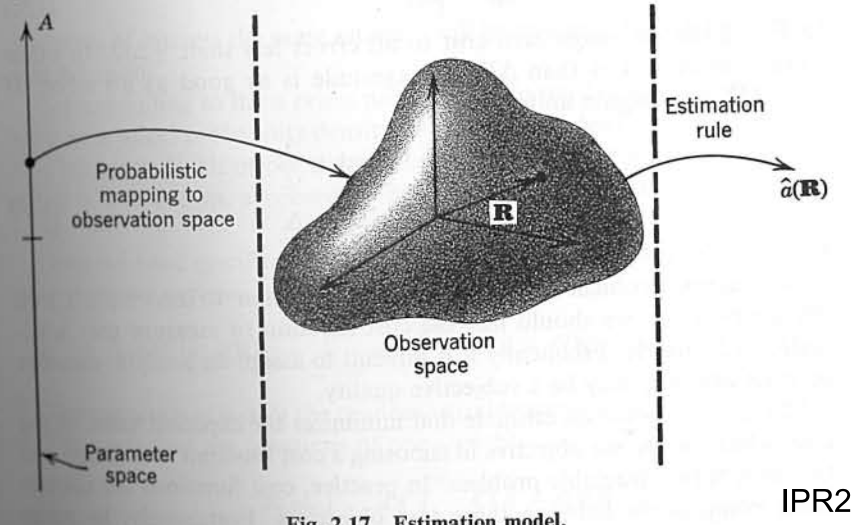


Fig. 2.17 Estimation model.

source models we encountered in the hypothesis-testing problem. To correspond with each of these models of the parameter space, we shall develop suitable estimation rules. We start with the random parameter case.

2.4.1 Random Parameters: Bayes Estimation

In the Bayes detection problem we saw that the two quantities we had to specify were the set of costs C_{ij} and the a priori probabilities P_i. The cost matrix assigned a cost to each possible course of action. Because there were M hypotheses and M possible decisions, there were M^2 costs. In the estimation problem a and $\hat{a}(\mathbf{R})$ are continuous variables. Thus we must assign a cost to all pairs $[a, \hat{a}(\mathbf{R})]$ over the range of interest. This is a function of two variables which we denote as $C(a, \hat{a})$. In many cases of interest it is realistic to assume that the cost depends only on the error of the estimate. We define this error as

$$a_\epsilon(\mathbf{R}) \triangleq \hat{a}(\mathbf{R}) - a. \tag{118}$$

The cost function $C(a_\epsilon)$ is a function of a single variable. Some typical cost functions are shown in Fig. 2.18. In Fig. 2.18*a* the cost function is simply the square of the error:

$$C(a_\epsilon) = a_\epsilon^2. \tag{119}$$

This cost is commonly referred to as the squared error cost function. We see that it accentuates the effects of large errors. In Fig. 2.18*b* the cost function is the absolute value of the error:

$$C(a_\epsilon) = |a_\epsilon|. \tag{120}$$

In Fig. 2.18*c* we assign zero cost to all errors less than ±Δ/2. In other words, an error less than Δ/2 in magnitude is as good as no error. If $a_\epsilon > \Delta/2$, we assign a uniform value:

$$\begin{aligned} C(a_\epsilon) &= 0, \qquad |a_\epsilon| \le \frac{\Delta}{2}, \\ &= 1, \qquad |a_\epsilon| > \frac{\Delta}{2}. \end{aligned} \tag{121}$$

In a given problem we choose a cost function to accomplish two objectives. First, we should like the cost function to measure user satisfaction adequately. Frequently it is difficult to assign an analytic measure to what basically may be a subjective quality.

Our goal is to find an estimate that minimizes the expected value of the cost. Thus our second objective in choosing a cost function is to assign one that results in a tractable problem. In practice, cost functions are usually some compromise between these two objectives. Fortunately, in many problems of interest the same estimate will be optimum for a large class of cost functions.

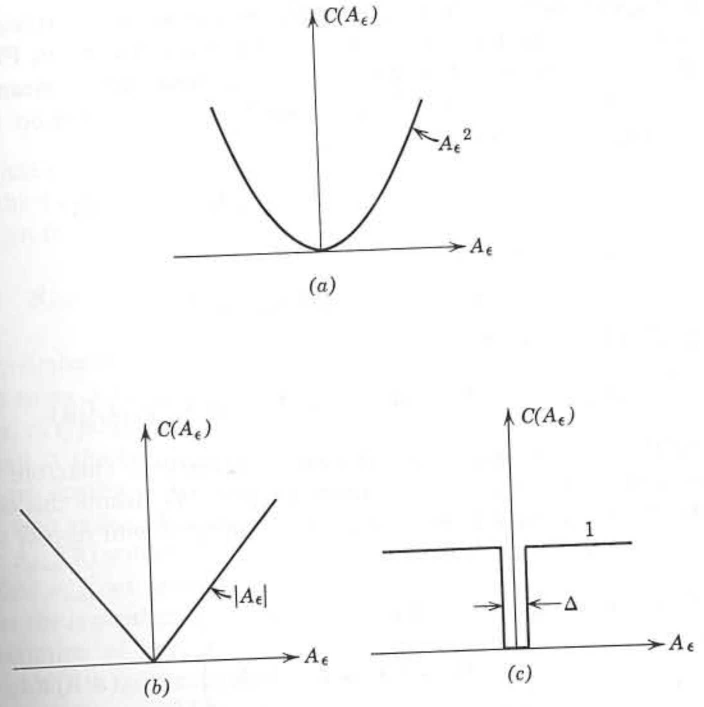


Fig. 2.18 Typical cost functions: (*a*) mean-square error; (*b*) absolute error; (*c*) uniform cost function.

Corresponding to the a priori probabilities in the detection problem, we have an a priori probability density $p_a(A)$ in the random parameter estimation problem. In all of our discussions we assume that $p_a(A)$ is known. If $p_a(A)$ is not known, a procedure analogous to the minimax test may be used.

Once we have specified the cost function and the a priori probability, we may write an expression for the risk:

$$\mathcal{R} \triangleq E\{C[a, \hat{a}(\mathbf{R})]\} = \int_{-\infty}^{\infty} dA \int_{-\infty}^{\infty} C[A - \hat{a}(\mathbf{R})]p_{a,\mathbf{r}}(A, \mathbf{R})\, d\mathbf{R}. \tag{122}$$

The expectation is over the random variable a and the observed variables $\mathbf{r}$. For costs that are functions of one variable only (122) becomes

$$\mathcal{R} = \int_{-\infty}^{\infty} dA \int_{-\infty}^{\infty} C[A - \hat{a}(\mathbf{R})]p_{a,\mathbf{r}}(A, \mathbf{R})\, d\mathbf{R}. \quad (123)$$

The Bayes estimate is the estimate that minimizes the risk. It is straightforward to find the Bayes estimates for the cost functions in Fig. 2.18. For the cost function in Fig. 2.18*a*, the risk corresponds to mean-square error. We denote the risk for the mean-square error criterion as $\mathcal{R}_{\text{ms}}$. Substituting (122) into (123), we have

$$\mathcal{R}_{\text{ms}} = \int_{-\infty}^{\infty} dA \int_{-\infty}^{\infty} d\mathbf{R}[A - \hat{a}(\mathbf{R})]^2 p_{a,\mathbf{r}}(A, \mathbf{R}). \tag{124}$$

The joint density can be rewritten as

$$p_{a,\mathbf{r}}(A, \mathbf{R}) = p_{\mathbf{r}}(\mathbf{R}) p_{a|\mathbf{r}}(A|\mathbf{R}). \tag{125}$$

Using (125) in (124), we have

$$\mathcal{R}_{\text{ms}} = \int_{-\infty}^{\infty} d\mathbf{R}\, p_{\mathbf{r}}(\mathbf{R}) \int_{-\infty}^{\infty} dA[A - \hat{a}(\mathbf{R})]^2 p_{a|\mathbf{r}}(A|\mathbf{R}). \tag{126}$$

Now the inner integral and $p_{\mathbf{r}}(\mathbf{R})$ are non-negative. Therefore we can minimize $\mathcal{R}_{\text{ms}}$ by minimizing the inner integral. We denote this estimate $\hat{a}_{\text{ms}}(\mathbf{R})$. To find it we differentiate the inner integral with respect to $\hat{a}(\mathbf{R})$ and set the result equal to zero:

$$\frac{d}{d\hat{a}} \int_{-\infty}^{\infty} dA[A - \hat{a}(\mathbf{R})]^2 p_{a|\mathbf{r}}(A|\mathbf{R})$$
$$= -2 \int_{-\infty}^{\infty} A p_{a|\mathbf{r}}(A|\mathbf{R})\, dA + 2\hat{a}(\mathbf{R}) \int_{-\infty}^{\infty} p_{a|\mathbf{r}}(A|\mathbf{R})\, dA. \tag{127}$$

Setting the result equal to zero and observing that the second integral equals 1, we have

$$\hat{a}_{\text{ms}}(\mathbf{R}) = \int_{-\infty}^{\infty} dA\, A p_{a|\mathbf{r}}(A|\mathbf{R}). \tag{128}$$

This is a unique minimum, for the second derivative equals two. The term on the right side of (128) is familiar as the mean of the a posteriori density (or the conditional mean).

Looking at (126), we see that if $\hat{a}(\mathbf{R})$ is the conditional mean the inner integral is just the a posteriori variance (or the conditional variance). Therefore the minimum value of $\mathcal{R}_{\text{ms}}$ is just the average of the conditional variance over all observations $\mathbf{R}$.

To find the Bayes estimate for the absolute value criterion in Fig. 2.18*b* we write

$$\mathcal{R}_{\text{abs}} = \int_{-\infty}^{\infty} d\mathbf{R}\, p_{\mathbf{r}}(\mathbf{R}) \int_{-\infty}^{\infty} dA[|A - \hat{a}(\mathbf{R})|] p_{a|\mathbf{r}}(A|\mathbf{R}). \tag{129}$$

To minimize the inner integral we write

$$I(\mathbf{R}) = \int_{-\infty}^{\hat{a}(\mathbf{R})} dA[\hat{a}(\mathbf{R}) - A]\, p_{a|\mathbf{r}}(A|\mathbf{R}) + \int_{\hat{a}(\mathbf{R})}^{\infty} dA[A - \hat{a}(\mathbf{R})]\, p_{a|\mathbf{r}}(A|\mathbf{R}). \tag{130}$$

Differentiating with respect to $\hat{a}(\mathbf{R})$ and setting the result equal to zero, we have

$$\int_{-\infty}^{\hat{a}_{\text{abs}}(\mathbf{R})} dA\, p_{a|\mathbf{r}}(A|\mathbf{R}) = \int_{\hat{a}_{\text{abs}}(\mathbf{R})}^{\infty} dA\, p_{a|\mathbf{r}}(A|\mathbf{R}). \tag{131}$$

This is just the definition of the median of the a posteriori density.

The third criterion is the uniform cost function in Fig. 2.18*c*. The risk expression follows easily:

$$\mathcal{R}_{\text{unf}} = \int_{-\infty}^{\infty} d\mathbf{R}\, p_{\mathbf{r}}(\mathbf{R}) \left[1 - \int_{\hat{a}_{\text{unf}}(\mathbf{R}) - \Delta/2}^{\hat{a}_{\text{unf}}(\mathbf{R}) + \Delta/2} p_{a|\mathbf{r}}(A|\mathbf{R})\, dA\right]. \tag{132}$$

To minimize this equation we maximize the inner integral. Of particular interest to us is the case in which Δ is an arbitrarily small but nonzero number. A typical a posteriori density is shown in Fig. 2.19. We see that for small Δ the best choice for $\hat{a}(\mathbf{R})$ is the value of A at which the a posteriori density has its maximum. We denote the estimate for this special case as $\hat{a}_{\text{map}}(\mathbf{R})$, the *maximum a posteriori* estimate. In the sequel we use $\hat{a}_{\text{map}}(\mathbf{R})$ without further reference to the uniform cost function.

To find $\hat{a}_{\text{map}}$ we must have the location of the maximum of $p_{a|\mathbf{r}}(A|\mathbf{R})$. Because the logarithm is a monotone function, we can find the location of the maximum of $\ln p_{a|\mathbf{r}}(A|\mathbf{R})$ equally well. As we saw in the detection problem, this is frequently more convenient.

If the maximum is interior to the allowable range of A and $\ln p_{a|\mathbf{r}}(A|\mathbf{R})$ has a continuous first derivative then a necessary, but not sufficient, condition for a maximum can be obtained by differentiating $\ln p_{a|\mathbf{r}}(A|\mathbf{R})$ with respect to A and setting the result equal to zero:

$$\left.\frac{\partial \ln p_{a|\mathbf{r}}(A|\mathbf{R})}{\partial A}\right|_{A=\hat{a}(\mathbf{R})} = 0. \tag{133}$$

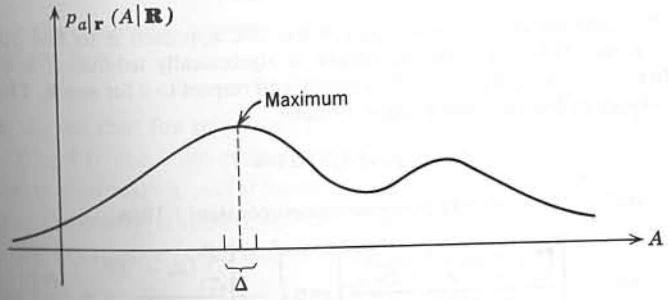


Fig. 2.19 An a posteriori density.

We refer to (133) as the MAP equation. In each case we must check to see if the solution is the absolute maximum.

We may rewrite the expression for $p_{a|\mathbf{r}}(A|\mathbf{R})$ to separate the role of the observed vector $\mathbf{R}$ and the a priori knowledge:

$$p_{a|\mathbf{r}}(A|\mathbf{R}) = \frac{p_{\mathbf{r}|a}(\mathbf{R}|A)p_a(A)}{p_{\mathbf{r}}(\mathbf{R})}. \tag{134}$$

Taking logarithms,

$$\ln p_{a|\mathbf{r}}(A|\mathbf{R}) = \ln p_{\mathbf{r}|a}(\mathbf{R}|A) + \ln p_a(A) - \ln p_{\mathbf{r}}(\mathbf{R}). \tag{135}$$

For MAP estimation we are interested only in finding the value of A where the left-hand side is maximum. Because the last term on the right-hand side is not a function of A, we can consider just the function

$$l(A) \triangleq \ln p_{\mathbf{r}|a}(\mathbf{R}|A) + \ln p_a(A). \tag{136}$$

The first term gives the probabilistic dependence of $\mathbf{R}$ on A and the second describes a priori knowledge.

The MAP equation can be written as

$$\left.\frac{\partial l(A)}{\partial A}\right|_{A=\hat{a}(\mathbf{R})} = \left.\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\right|_{A=\hat{a}(\mathbf{R})} + \left.\frac{\partial \ln p_a(A)}{\partial A}\right|_{A=\hat{a}(\mathbf{R})} = 0. \tag{137}$$

Our discussion in the remainder of the book emphasizes minimum mean-square error and maximum a posteriori estimates.

To study the implications of these two estimation procedures we consider several examples.

Example 2. Let

$$r_i = a + n_i, \qquad i = 1, 2, \ldots, N. \tag{138}$$

We assume that a is Gaussian, $N(0, \sigma_a)$, and that the n_i are each independent Gaussian variables $N(0, \sigma_n)$. Then

$$p_{\mathbf{r}|a}(\mathbf{R}|A) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_n} \exp\left(-\frac{(R_i - A)^2}{2\sigma_n^2}\right),$$
$$p_a(A) = \frac{1}{\sqrt{2\pi}\,\sigma_a} \exp\left(-\frac{A^2}{2\sigma_a^2}\right). \tag{139}$$

To find $\hat{a}_{ms}(\mathbf{R})$ we need to know $p_{a|\mathbf{r}}(A|\mathbf{R})$. One approach is to find $p_{\mathbf{r}}(\mathbf{R})$ and substitute it into (134), but this procedure is algebraically tedious. It is easier to observe that $p_{a|\mathbf{r}}(A|\mathbf{R})$ is a probability density with respect to a for any $\mathbf{R}$. Thus $p_{\mathbf{r}}(\mathbf{R})$ just contributes to the constant needed to make

$$\int_{-\infty}^{\infty} p_{a|\mathbf{r}}(A|\mathbf{R})\, dA = 1. \tag{140}$$

(In other words, $p_{\mathbf{r}}(\mathbf{R})$ is simply a normalization constant.) Thus

$$p_{a|\mathbf{r}}(A|\mathbf{R}) = \left[\frac{\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_n}\right)\frac{1}{\sqrt{2\pi}\,\sigma_a}}{p_{\mathbf{r}}(\mathbf{R})}\right] \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^{N}(R_i - A)^2}{\sigma_n^2} + \frac{A^2}{\sigma_a^2}\right]\right\}. \tag{141}$$

Rearranging the exponent, completing the square, and absorbing terms depending only on R_i^2 into the constant, we have

$$p_{a|\mathbf{r}}(A|\mathbf{R}) = k(\mathbf{R}) \exp\left\{-\frac{1}{2\sigma_p^2}\left[A -- \frac{\sigma_a^2}{\sigma_a^2 + \sigma_n^2/N}\left(\frac{1}{N}\sum_{i=1}^{N} R_i\right)\right]^2\right\}, \tag{142}$$

where

$$\sigma_p^2 \triangleq \left(\frac{1}{\sigma_a^2} + \frac{N}{\sigma_n^2}\right)^{-1} = \frac{\sigma_a^2\sigma_n^2}{N\sigma_a^2 + \sigma_n^2} \tag{143}$$

is the a posteriori variance.

We see that $p_{a|\mathbf{r}}(A|\mathbf{R})$ is just a Gaussian density. The estimate $\hat{a}_{ms}(\mathbf{R})$ is just the conditional mean

$$\hat{a}_{ms}(\mathbf{R}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_n^2/N}\left(\frac{1}{N}\sum_{i=1}^{N} R_i\right). \tag{144}$$

Because the a posteriori variance is not a function of $\mathbf{R}$, the mean-square risk equals the a posteriori variance (see (126)).

Two observations are useful:

1. The R_i enter into the a posteriori density only through their sum. Thus

$$l(\mathbf{R}) = \sum_{i=1}^{N} R_i \tag{145}$$

is a *sufficient statistic*. This idea of a sufficient statistic is identical to that in the detection problem.

2. The estimation rule uses the information available in an intuitively logical manner. If $\sigma_a^2 \ll \sigma_n^2/N$, the a priori knowledge is much better than the observed data and the estimate is very close to the a priori mean. (In this case, the a priori mean is zero.) On the other hand, if $\sigma_a^2 \gg \sigma_n^2/N$, the a priori knowledge is of little value and the estimate uses primarily the received data. In the limit $\hat{a}_{ms}$ is just the arithmetic average of the R_i.

$$\lim_{\frac{\sigma_n^2}{N\sigma_a^2} \to 0} \hat{a}_{ms}(\mathbf{R}) = \frac{1}{N}\sum_{i=1}^{N} R_i. \tag{146}$$

The MAP estimate for this case follows easily. Looking at (142), we see that because the density is Gaussian the maximum value of $p_{a|\mathbf{r}}(A|\mathbf{R})$ occurs at the conditional mean. Thus

$$\hat{a}_{map}(\mathbf{R}) = \hat{a}_{ms}(\mathbf{R}). \tag{147}$$

Because the conditional median of a Gaussian density occurs at the conditional mean, we also have

$$\hat{a}_{abs}(\mathbf{R}) = \hat{a}_{ms}(\mathbf{R}). \tag{148}$$

Thus we see that for this particular example all three cost functions in Fig. 2.18 lead to the same estimate. This invariance to the choice of a cost function is obviously a useful feature because of the subjective judgments that are frequently involved in choosing $C(a_\epsilon)$. Some conditions under which this invariance holds are developed in the next two properties.†

† These properties are due to Sherman [20]. Our derivation is similar to that given by Viterbi [36].

Property 1. We assume that the cost function $C(a_\epsilon)$ is a symmetric, convex-upward function and that the a posteriori density $p_{a|\mathbf{r}}(A|\mathbf{R})$ is symmetric about its conditional mean; that is,

$$C(a_\epsilon) = C(-a_\epsilon) \qquad \text{(symmetry)}, \quad (149)$$

$$C(bx_1 + (1 - b)x_2) \leq bC(x_1) + (1 - b)\, C(x_2) \qquad \text{(convexity)} \quad (150)$$

for any b inside the range (0, 1) and for all x_1 and x_2. Equation 150 simply says that all chords lie above or on the cost function.

This condition is shown in Fig. 2.20*a*. If the inequality is strict whenever $x_1 \neq x_2$, we say the cost function is strictly convex (upward). Defining

$$z \triangleq a - \hat{a}_{\text{ms}} = a - E[a|\mathbf{R}] \qquad (151)$$

the symmetry of the a posteriori density implies

$$p_{z|\mathbf{r}}(Z|\mathbf{R}) = p_{z|\mathbf{r}}(-Z|\mathbf{R}). \qquad (152)$$

The estimate $\hat{a}$ that minimizes any cost function in this class is identical to $\hat{a}_{\text{ms}}$ (which is the conditional mean).

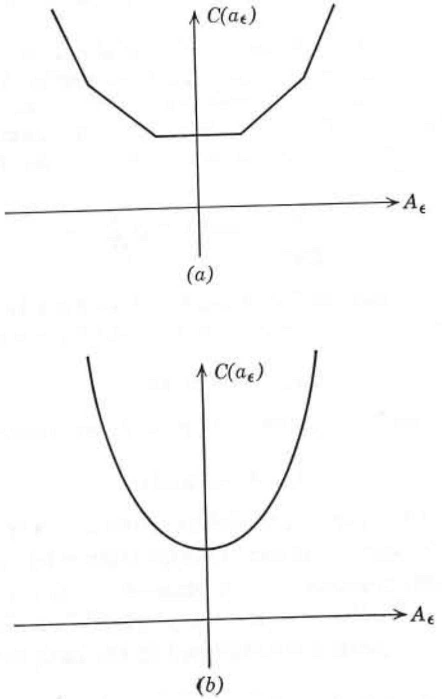


Fig. 2.20 Symmetric convex cost functions: (*a*) convex; (*b*) strictly convex.

Proof. As before we can minimize the conditional risk [see (126)]. Define

$$\mathcal{R}_B(\hat{a}|\mathbf{R}) \triangleq E_a[C(\hat{a} - a)|\mathbf{R}] = E_a[C(a - \hat{a})|\mathbf{R}], \qquad (153)$$

where the second equality follows from (149). We now write four equivalent expressions for $\mathcal{R}_B(\hat{a}|\mathbf{R})$:

$$\mathcal{R}_B(\hat{a}|\mathbf{R}) = \int_{-\infty}^{\infty} C(\hat{a} - \hat{a}_{\text{ms}} - Z) p_{z|\mathbf{r}}(Z|\mathbf{R})\, dZ \qquad (154)$$

[Use (151) in (153)]

$$= \int_{-\infty}^{\infty} C(\hat{a} - \hat{a}_{\text{ms}} + Z) p_{z|\mathbf{r}}(Z|\mathbf{R})\, dZ \qquad (155)$$

[(152) implies this equality]

$$= \int_{-\infty}^{\infty} C(\hat{a}_{\text{ms}} - \hat{a} - Z) p_{z|\mathbf{r}}(Z|\mathbf{R})\, dZ \qquad (156)$$

[(149) implies this equality]

$$= \int_{-\infty}^{\infty} C(\hat{a}_{\text{ms}} - \hat{a} + Z) p_{z|\mathbf{r}}(Z|\mathbf{R})\, dZ \qquad (157)$$

[(152) implies this equality].

We now use the convexity condition (150) with the terms in (155) and (157):

$$\begin{aligned}\mathcal{R}_B(\hat{a}|\mathbf{R}) &= \tfrac{1}{2}E(\{C[Z + (\hat{a}_{\text{ms}} - \hat{a})] + C[Z - (\hat{a}_{\text{ms}} - \hat{a})]\}|\mathbf{R}) \\ &\geq E\{C[\tfrac{1}{2}(Z + (\hat{a}_{\text{ms}} - \hat{a})) + \tfrac{1}{2}(Z - (\hat{a}_{\text{ms}} - \hat{a}))]|\mathbf{R}\} \\ &= E[C(Z)|\mathbf{R}]. \end{aligned} \qquad (158)$$

Equality will be achieved in (158) if $\hat{a}_{\text{ms}} = \hat{a}$. This completes the proof. If $C(a_\epsilon)$ is strictly convex, we will have the additional result that the minimizing estimate $\hat{a}$ is unique and equals $\hat{a}_{\text{ms}}$.

To include cost functions like the uniform cost functions which are not convex we need a second property.

Property 2. We assume that the cost function is a symmetric, nondecreasing function and that the a posteriori density $p_{a|\mathbf{r}}(A|\mathbf{R})$ is a symmetric (about the conditional mean), unimodal function that satisfies the condition

$$\lim_{x \to \infty} C(x) p_{a|\mathbf{r}}(x|\mathbf{R}) = 0.$$

The estimate $\hat{a}$ that minimizes any cost function in this class is identical to $\hat{a}_{\text{ms}}$. The proof of this property is similar to the above proof [36].

The significance of these two properties should not be underemphasized. Throughout the book we consider only minimum mean-square and maximum a posteriori probability estimators. Properties 1 and 2 ensure that whenever the a posteriori densities satisfy the assumptions given above the estimates that we obtain will be optimum for a large class of cost functions. Clearly, if the a posteriori density is Gaussian, it will satisfy the above assumptions.

We now consider two examples of a different type.

Example 3. The variable a appears in the signal in a nonlinear manner. We denote this dependence by $s(A)$. Each observation r_i consists of $s(A)$ plus a Gaussian random variable n_i, $N(0, \sigma_n)$. The n_i are statistically independent of each other and a. Thus

$$r_i = s(A) + n_i. \tag{159}$$

Therefore

$$p_{a|\mathbf{r}}(A|\mathbf{R}) = k(\mathbf{R}) \exp\left(-\frac{1}{2}\left\{\frac{\sum_{i=1}^{N} [R_i - s(A)]^2}{\sigma_n{}^2} + \frac{A^2}{\sigma_a{}^2}\right\}\right). \tag{160}$$

This expression cannot be further simplified without specifying $s(A)$ explicitly.

The MAP equation is obtained by substituting (160) into (137)

$$\hat{a}_{\text{map}}(\mathbf{R}) = \frac{\sigma_a{}^2}{\sigma_n{}^2} \sum_{i=1}^{N} [R_i - s(A)] \left.\frac{\partial s(A)}{\partial A}\right|_{A = \hat{a}_{\text{map}}(\mathbf{R})}. \tag{161}$$

To solve this explicitly we must specify $s(A)$. We shall find that an analytic solution is generally not possible when $s(A)$ is a nonlinear function of A.

Another type of problem that frequently arises is the estimation of a parameter in a probability density.

Example 4. The number of events in an experiment obey a Poisson law with mean value a. Thus

$$\Pr(n \text{ events} \mid a = A) = \frac{A^n}{n!} \exp(-A), \qquad n = 0, 1, \ldots. \tag{162}$$

We want to observe the number of events and estimate the parameter a of the Poisson law. We shall assume that a is a random variable with an exponential density

$$p_a(A) = \begin{cases} \lambda \exp(-\lambda A), & A > 0, \\ 0, & \text{elsewhere.} \end{cases} \tag{163}$$

The a posteriori density of a is

$$p_{a|n}(A|N) = \frac{\Pr(n = N \mid a = A) p_a(A)}{\Pr(n = N)}. \tag{164}$$

Substituting (162) and (163) into (164), we have

$$p_{a|n}(A|N) = k(N)[A^N \exp(-A(1 + \lambda))], \qquad A \geq 0, \tag{165}$$

where

$$k(N) = \frac{(1 + \lambda)^{N+1}}{N!} \tag{166}$$

in order for the density to integrate to 1. (As already pointed out, the constant is unimportant for MAP estimation but is needed if we find the MS estimate by integrating over the conditional density.)

The mean-square estimate is the conditional mean:

$$\begin{aligned} \hat{a}_{\text{ms}}(N) &= \frac{(1 + \lambda)^{N+1}}{N!} \int_0^\infty A^{N+1} \exp[-A(1 + \lambda)]\, dA \\ &= \frac{(1 + \lambda)^{N+1}}{(1 + \lambda)^{N+2}} (N + 1) = \left(\frac{1}{\lambda + 1}\right)(N + 1). \end{aligned} \tag{167}$$

To find $\hat{a}_{\text{map}}$ we take the logarithm of (165)

$$\ln p_{a|n}(A|N) = N \ln A - A(1 + \lambda) + \ln k(N). \tag{168}$$

By differentiating with respect to A, setting the result equal to zero, and solving, we obtain

$$\hat{a}_{\text{map}}(N) = \frac{N}{1 + \lambda}. \tag{169}$$

Observe that $\hat{a}_{\text{map}}$ is not equal to $\hat{a}_{\text{ms}}$.

Other examples are developed in the problems. The principal results of this section are the following:

1. The minimum mean-square error estimate (MMSE) is always the mean of the a posteriori density (the conditional mean).
2. The maximum a posteriori estimate (MAP) is the value of A at which the a posteriori density has its maximum.
3. For a large class of cost functions the optimum estimate is the conditional mean whenever the a posteriori density is a unimodal function which is symmetric about the conditional mean.

These results are the basis of most of our estimation work. As we study more complicated problems, the only difficulty we shall encounter is the actual evaluation of the conditional mean or maximum. In many cases o f interest the MAP and MMSE estimates will turn out to be equal.

We now turn to the second class of estimation problems described in the introduction.

2.4.2 Real (Nonrandom) Parameter Estimation†

In many cases it is unrealistic to treat the unknown parameter as a random variable. The problem formulation on pp. 52–53 is still appropriate. Now, however, the parameter is assumed to be nonrandom, and we want to design an estimation procedure that is good in some sense.

† The beginnings of classical estimation theory can be attributed to Fisher [5, 6, 7, 8]. Many discussions of the basic ideas are now available (e.g., Cramér [9]), Wilks [10], or Kendall and Stuart [11]).

A logical first approach is to try to modify the Bayes procedure in the last section to eliminate the average over $p_a(A)$. As an example, consider a mean-square error criterion,

$$\mathcal{R}(A) \triangleq \int_{-\infty}^{\infty} [\hat{a}(\mathbf{R}) - A]^2\, p_{\mathbf{r}|a}(\mathbf{R}|A)\, d\mathbf{R}, \qquad (170)$$

where the expectation is only over $\mathbf{R}$, for it is the only random variable in the model. Minimizing $\mathcal{R}(A)$, we obtain

$$\hat{a}_{\text{ms}}(\mathbf{R}) = A. \qquad (171)$$

The answer is correct, but not of any value, for A is the unknown quantity that we are trying to find. Thus we see that this direct approach is not fruitful. A more useful method in the nonrandom parameter case is to examine other possible measures of quality of estimation procedures and then to see whether we can find estimates that are good in terms of these measures.

The first measure of quality to be considered is the expectation of the estimate

$$E[\hat{a}(\mathbf{R})] \triangleq \int_{-\infty}^{+\infty} \hat{a}(\mathbf{R})\, p_{\mathbf{r}|a}(\mathbf{R}|A)\, d\mathbf{R}. \qquad (172)$$

The possible values of the expectation can be grouped into three classes

1. If $E[\hat{a}(\mathbf{R})] = A$, for all values of A, we say that the estimate is *unbiased*. This statement means that the average value of the estimates equals the quantity we are trying to estimate.
2. If $E[\hat{a}(\mathbf{R})] = A + B$, where B is not a function of A, we say that the estimate has a *known bias*. We can always obtain an unbiased estimate by subtracting B from $\hat{a}(\mathbf{R})$.
3. If $E[\hat{a}(\mathbf{R})] = A + B(A)$, we say that the estimate has an *unknown bias*. Because the bias depends on the unknown parameter, we cannot simply subtract it out.

Clearly, even an unbiased estimate may give a bad result on a particular trial. A simple example is shown in Fig. 2.21. The probability density of the estimate is centered around A, but the variance of this density is large enough that big errors are probable.

A second measure of quality is the variance of estimation error:

$$\text{Var}\,[\hat{a}(\mathbf{R}) - A] = E\{[\hat{a}(\mathbf{R}) - A]^2\} - B^2(A). \qquad (173)$$

This provides a measure of the spread of the error. In general, we shall try to find unbiased estimates with small variances. There is no straightforward minimization procedure that will lead us to the minimum variance unbiased estimate. Therefore we are forced to try an estimation procedure to see how well it works.

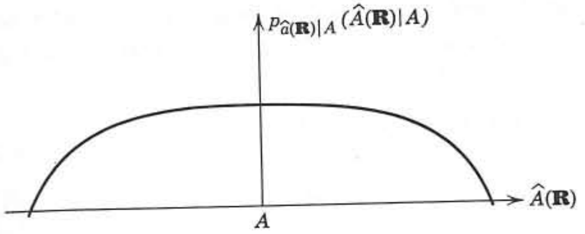


Fig. 2.21 Probability density for an estimate.

Maximum Likelihood Estimation. There are several ways to motivate the estimation procedure that we shall use. Consider the simple estimation problem outlined in Example 1. Recall that

$$r = A + n, \qquad (174)$$

$$p_{r|a}(R|A) = (\sqrt{2\pi}\,\sigma_n)^{-1} \exp\,[-\tfrac{1}{2}(R - A)^2]. \qquad (175)$$

We choose as our estimate the value of A that most likely caused a given value of R to occur. In this simple additive case we see that this is the same as choosing the most probable value of the noise ($N = 0$) and subtracting it from R. We denote the value obtained by using this procedure as a maximum likelihood estimate.

$$\hat{a}_{ml}(R) = R. \qquad (176)$$

In the general case we denote the function $p_{\mathbf{r}|a}(\mathbf{R}|A)$, viewed as a function of A, as the *likelihood function*. Frequently we work with the logarithm, $\ln p_{\mathbf{r}|a}(\mathbf{R}|A)$, and denote it as the *log likelihood function*. The maximum likelihood estimate $\hat{a}_{ml}(\mathbf{R})$ is that value of A at which the likelihood function is a maximum. If the maximum is interior to the range of A, and $\ln p_{\mathbf{r}|a}(\mathbf{R}|A)$ has a continuous first derivative, then a necessary condition on $\hat{a}_{ml}(\mathbf{R})$ is obtained by differentiating $\ln p_{\mathbf{r}|a}(\mathbf{R}|A)$ with respect to A and setting the result equal to zero:

$$\left.\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\right|_{A=\hat{a}_{ml}(\mathbf{R})} = 0. \qquad (177)$$

This equation is called the *likelihood equation*. Comparing (137) and (177), we see that the ML estimate corresponds mathematically to the limiting case of a MAP estimate in which the a priori knowledge approaches zero.

In order to see how effective the ML procedure is we can compute the bias and the variance. Frequently this is difficult to do. Rather than approach the problem directly, we shall first derive a lower bound on the variance on *any* unbiased estimate. Then we shall see how the variance of $\hat{a}_{ml}(\mathbf{R})$ compares with this lower bound.

Cramér-Rao Inequality: Nonrandom Parameters. We now want to consider the variance of *any* estimate $\hat{a}(\mathbf{R})$ of the real variable A. We shall prove the following statement.

Theorem. (a) If $\hat{a}(\mathbf{R})$ is *any* unbiased estimate of A, then

$$\operatorname{Var}[\hat{a}(\mathbf{R}) - A] \geq \left(E\left\{\left[\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\right]^2\right\}\right)^{-1} \tag{178}$$

or, equivalently,

(b)

$$\operatorname{Var}[\hat{a}(\mathbf{R}) - A] \geq \left\{-E\left[\frac{\partial^2 \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^2}\right]\right\}^{-1}, \tag{179}$$

where the following conditions are assumed to be satisfied:

(c)

$$\frac{\partial p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A} \quad \text{and} \quad \frac{\partial^2 p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^2}$$

exist and are absolutely integrable.

The inequalities were first stated by Fisher [6] and proved by Dugué [31]. They were also derived by Cramér [9] and Rao [12] and are usually referred to as the Cramér-Rao bound. Any estimate that satisfies the bound with an equality is called an *efficient* estimate.

The proof is a simple application of the Schwarz inequality. Because $\hat{a}(\mathbf{R})$ is unbiased,

$$E[\hat{a}(\mathbf{R}) - A] \triangleq \int_{-\infty}^{\infty} p_{\mathbf{r}|a}(\mathbf{R}|A)[\hat{a}(\mathbf{R}) - A]\, d\mathbf{R} = 0. \tag{180}$$

Differentiating both sides with respect to A, we have

$$\frac{d}{dA}\int_{-\infty}^{\infty} p_{\mathbf{r}|a}(\mathbf{R}|A)[\hat{a}(\mathbf{R}) - A]\, d\mathbf{R} = \int_{-\infty}^{\infty} \frac{\partial}{\partial A}\{p_{\mathbf{r}|a}(\mathbf{R}|A)[\hat{a}(\mathbf{R}) - A]\}\, d\mathbf{R} = 0, \tag{181}$$

where condition (c) allows us to bring the differentiation inside the integral. Then

$$-\int_{-\infty}^{\infty} p_{\mathbf{r}|a}(\mathbf{R}|A)\, d\mathbf{R} + \int_{-\infty}^{\infty} \frac{\partial p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}[\hat{a}(\mathbf{R}) - A]\, d\mathbf{R} = 0. \tag{182}$$

The first integral is just $+1$. Now observe that

$$\frac{\partial p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A} = \frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\, p_{\mathbf{r}|a}(\mathbf{R}|A). \tag{183}$$

Substituting (183) into (182), we have

$$\int_{-\infty}^{\infty} \frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\, p_{\mathbf{r}|a}(\mathbf{R}|A)[\hat{a}(\mathbf{R}) - A]\, d\mathbf{R} = 1. \tag{184}$$

Rewriting, we have

$$\int_{-\infty}^{\infty} \left[\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\sqrt{p_{\mathbf{r}|a}(\mathbf{R}|A)}\right]\left[\sqrt{p_{\mathbf{r}|a}(\mathbf{R}|A)}\,[\hat{a}(\mathbf{R}) - A]\right] d\mathbf{R} = 1, \tag{185}$$

and, using the Schwarz inequality, we have

$$\left\{\int_{-\infty}^{\infty} \left[\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\right]^2 p_{\mathbf{r}|a}(\mathbf{R}|A)\, d\mathbf{R}\right\} \times \left\{\int_{-\infty}^{\infty} [\hat{a}(\mathbf{R}) - A]^2\, p_{\mathbf{r}|a}(\mathbf{R}|A)\, d\mathbf{R}\right\} \geq 1, \tag{186}$$

where we recall from the derivation of the Schwarz inequality that equality holds if and only if

$$\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A} = [\hat{a}(\mathbf{R}) - A]\, k(A), \tag{187}$$

for all $\mathbf{R}$ and A. We see that the two terms of the left side of (186) are the expectations in statement (*a*) of (178). Thus,

$$E\{[\hat{a}(\mathbf{R}) - A]^2\} \geq \left\{E\left[\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\right]^2\right\}^{-1}. \tag{188}$$

To prove statement (b) we observe

$$\int_{-\infty}^{\infty} p_{\mathbf{r}|a}(\mathbf{R}|A)\, d\mathbf{R} = 1. \tag{189}$$

Differentiating with respect to A, we have

$$\int_{-\infty}^{\infty} \frac{\partial p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\, d\mathbf{R} = \int_{-\infty}^{\infty} \frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\, p_{\mathbf{r}|a}(\mathbf{R}|A)\, d\mathbf{R} = 0. \tag{190}$$

Differentiating again with respect to A and applying (183), we obtain

$$\int_{-\infty}^{\infty} \frac{\partial^2 \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^2}\, p_{\mathbf{r}|a}(\mathbf{R}|A)\, d\mathbf{R} + \int_{-\infty}^{\infty} \left(\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\right)^2 p_{\mathbf{r}|a}(\mathbf{R}|A)\, d\mathbf{R} = 0 \tag{191}$$

or

$$E\left[\frac{\partial^2 \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^2}\right] = -E\left[\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\right]^2, \tag{192}$$

which together with (188) gives condition (b).

Several important observations should be made about this result.

1. It shows that any unbiased estimate must have a variance greater than a certain number.

2. If (187) is satisfied, the estimate $\hat{a}_{ml}(\mathbf{R})$ will satisfy the bound with an equality. We show this by combining (187) and (177). The left equality is the maximum likelihood equation. The right equality is (187):

$$0 = \frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\bigg|_{A=\hat{a}_{ml}(\mathbf{R})} = (\hat{a}(\mathbf{R}) - A)\, k(A)\bigg|_{A=\hat{a}_{ml}(\mathbf{R})}. \tag{193}$$

In order for the right-hand side to equal zero either

$$\hat{a}(\mathbf{R}) = \hat{a}_{ml}(\mathbf{R}) \tag{194}$$

or

$$k(\hat{a}_{ml}) = 0. \tag{195}$$

Because we want a solution that depends on the data, we eliminate (195) and require (194) to hold.

Thus, *if* an efficient estimate exists, it is $\hat{a}_{ml}(\mathbf{R})$ and can be obtained as a unique solution to the likelihood equation.

3. If an efficient estimate *does not* exist [i.e., $\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)/\partial A$ cannot be put into the form of (187)], we do not know how good $\hat{a}_{ml}(\mathbf{R})$ is. Further, we do not know how close the variance of any estimate will approach the bound.

4. In order to use the bound, we must verify that the estimate of concern is unbiased. Similar bounds can be derived simply for biased estimates (Problem 2.4.17).

We can illustrate the application of ML estimation and the Cramér–Rao inequality by considering Examples 2, 3, and 4. The observation model is identical. We now assume, however, that the parameters to be estimated are nonrandom variables.

Example 2. From (138) we have

$$r_i = A + n_i, \qquad i = 1, 2, \ldots, N. \tag{196}$$

Taking the logarithm of (139) and differentiating, we have

$$\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A} = \frac{N}{\sigma_n^2}\left(\frac{1}{N}\sum_{i=1}^{N} R_i - A\right). \tag{197}$$

Thus

$$\hat{a}_{ml}(\mathbf{R}) = \frac{1}{N}\sum_{i=1}^{N} R_i. \tag{198}$$

To find the bias we take the expectation of both sides,

$$E[\hat{a}_{ml}(\mathbf{R})] = \frac{1}{N}\sum_{i=1}^{N} E(R_i) = \frac{1}{N}\sum_{i=1}^{N} A = A, \tag{199}$$

so that $\hat{a}_{ml}(\mathbf{R})$ is unbiased.

Because the expression in (197) has the form required by (187), we know that $\hat{a}_{ml}(\mathbf{R})$ is an efficient estimate. To evaluate the variance we differentiate (197):

$$\frac{\partial^2 \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^2} = -\frac{N}{\sigma_n^2}. \tag{200}$$

Using (179) and the efficiency result, we have

$$\text{Var}\,[\hat{a}_{ml}(\mathbf{R}) - A] = \frac{\sigma_n^2}{N}. \tag{201}$$

Skipping Example 3 for the moment, we go to Example 4.

Example 4. Differentiating the logarithm of (162), we have

$$\begin{aligned}\frac{\partial \ln \Pr\,(n = N|A)}{\partial A} &= \frac{\partial}{\partial A}(N \ln A - A - \ln N!)\\ &= \frac{N}{A} - 1 = \frac{1}{A}(N - A).\end{aligned} \tag{202}$$

The ML estimate is

$$\hat{a}_{ml}(N) = N. \tag{203}$$

It is clearly unbiased and efficient. To obtain the variance we differentiate (202):

$$\frac{\partial^2 \ln \Pr\,(n = N|A)}{\partial A^2} = -\frac{N}{A^2}. \tag{204}$$

Thus

$$\text{Var}\,[\hat{a}_{ml}(N) - A] = \frac{A^2}{E(N)} = \frac{A^2}{A} = A. \tag{205}$$

In both Examples 2 and 4 we see that the ML estimates could have been obtained from the MAP estimates [let $\sigma_a \to \infty$ in (144) and recall that $\hat{a}_{ms}(\mathbf{R}) = \hat{a}_{map}(\mathbf{R})$ and let $\lambda \to 0$ in (169)].

We now return to Example 3.

Example 3. From the first term in the exponent in (160), we have

$$\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A} = \frac{1}{\sigma_n^2}\sum_{i=1}^{N} [R_i - s(A)]\frac{\partial s(A)}{\partial A}. \tag{206}$$

In general, the right-hand side cannot be written in the form required by (187), and therefore an unbiased efficient estimate does not exist.

The likelihood equation is

$$\left[\frac{\partial s(A)}{\partial A}\frac{1}{\sigma_n^2}\right]\left[\frac{1}{N}\sum_{i=1}^{N} R_i - s(A)\right]\bigg|_{A=\hat{a}_{ml}(\mathbf{R})} = 0. \tag{207}$$

If the range of $s(A)$ includes $(1/N)\sum_{i=1}^{N} R_i$, a solution exists:

$$s[\hat{a}_{ml}(\mathbf{R})] = \frac{1}{N}\sum_{i=1}^{N} R_i. \tag{208}$$

If (208) can be satisfied, then

$$\hat{a}_{ml}(\mathbf{R}) = s^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} R_i\right). \tag{209}$$

[Observe that (209) tacitly assumes that $s^{-1}(\cdot)$ exists. If it does not, then even in the absence of noise we shall be unable to determine A unambiguously. If we were designing a system, we would always choose an $s(\cdot)$ that allows us to find A unambiguously in the absence of noise.] If the range of $s(a)$ does not include $(1/N)\sum_{i=1}^{N} R_i$, the maximum is at an end point of the range.

We see that the maximum likelihood estimate commutes over nonlinear operations. (This is *not* true for MS or MAP estimation.) If it is unbiased, we evaluate the bound on the variance by differentiating (206):

$$\frac{\partial^2 \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^2} = \frac{1}{\sigma_n^2}\sum_{i=1}^{N}[R_i - s(A)]\frac{\partial^2 s(A)}{\partial A^2} - \frac{N}{\sigma_n^2}\left[\frac{\partial s(A)}{\partial A}\right]^2. \tag{210}$$

Observing that

$$E[r_i - s(A)] = E(n_i) = 0, \tag{211}$$

we obtain the following bound for any unbiased estimate,

$$\text{Var}\,[\hat{a}(\mathbf{R}) - A] \geq \frac{\sigma_n^2}{N[\partial s(A)/\partial A]^2}. \tag{212}$$

We see that the bound is exactly the same as that in Example 2 except for a factor $[\partial s(A)/\partial A]^2$. The intuitive reason for this factor and also some feeling for the conditions under which the bound will be useful may be obtained by inspecting the typical function shown in Fig. 2.22. Define

$$Y = s(A). \tag{213}$$

Then

$$r_i = Y + n_i. \tag{214}$$

The variance in estimating Y is just σ_n^2/N. However, if y_ϵ, the error in estimating Y, is small enough so that the slope is constant, then

$$A_\epsilon \simeq \frac{Y_\epsilon}{\left.\dfrac{\partial s(A)}{\partial A}\right|_{A=\hat{a}(\mathbf{R})}} \tag{215}$$

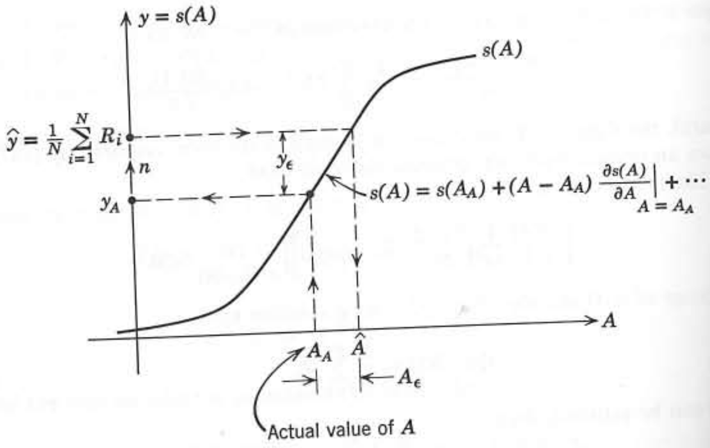

$y = s(A)$
$s(A)$
$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} R_i$
n
y_ϵ
y_A
$s(A) = s(A_A) + (A - A_A)\left.\frac{\partial s(A)}{\partial A}\right|_{A=A_A} + \cdots$
A
A_A
$\hat{A}$
A_ϵ
Actual value of A

Fig. 2.22 Behavior of error variance in the presence of small errors.

and

$$\text{Var}\,(a_\epsilon) \cong \frac{\text{Var}\,(y_\epsilon)}{[\partial s(A)/\partial A]^2} = \frac{\sigma_n^2}{N[\partial s(A)/\partial A]^2}. \tag{216}$$

We observe that if y_ϵ is large there will no longer be a simple linear relation between y_ϵ and a_ϵ. This tells us when we can expect the Cramér-Rao bound to give an accurate answer in the case in which the parameter enters the problem in a nonlinear manner. Specifically, whenever the estimation error is small, relative to $A\,\partial^2 s(A)/\partial A^2$, we should expect the actual variance to be close to the variance bound given by the Cramér-Rao inequality.

The properties of the ML estimate which are valid when the error is small are generally referred to as asymptotic. One procedure for developing them formally is to study the behavior of the estimate as the number of independent observations N approaches infinity. Under reasonably general conditions the following may be proved (e.g., Cramér [9], pp. 500–504).

1. The solution of the likelihood equation (177) converges in probability to the correct value of A as $N \to \infty$. *Any* estimate with this property is called consistent. Thus the ML estimate is consistent.
2. The ML estimate is asymptotically efficient; that is,

$$\lim_{N\to\infty} \frac{\text{Var}\,[\hat{a}_{ml}(\mathbf{R}) - A]}{\left(-E\left[\dfrac{\partial^2 \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^2}\right]\right)^{-1}} = 1.$$

3. The ML estimate is asymptotically Gaussian, $N(A, \sigma_{a_\epsilon})$.

These properties all deal with the behavior of ML estimates for large N. They provide some motivation for using the ML estimate even when an efficient estimate does not exist.

At this point a logical question is: "Do better estimation procedures than the maximum likelihood procedure exist?" Certainly if an efficient estimate does not exist, there may be unbiased estimates with lower variances. The difficulty is that there is no general rule for finding them. In a particular situation we can try to improve on the ML estimate. In almost all cases, however, the resulting estimation rule is more complex, and therefore we emphasize the maximum likelihood technique in all of our work with real variables.

A second logical question is: "Do better lower bounds than the Cramér–Rao inequality exist?" One straightforward but computationally tedious procedure is the Bhattacharyya bound. The Cramér–Rao bound uses $\partial^2 p_{\mathbf{r}|a}(\mathbf{R}|A)/\partial A^2$. Whenever an efficient estimate does not exist, a larger bound which involves the higher partial derivatives can be obtained. Simple derivations are given in [13] and [14] and in Problems 2.4.23–24. For the cases of interest to us the computation is too involved to make the bound of much practical value. A second bound is the Barankin bound (e.g. [15]). Its two major advantages are that it does not require the probability density to be differentiable and it gives the greatest lower bound. Its disadvantages are that it requires a maximization over a function to obtain the bound and the procedure for finding this maximum is usually not straightforward. Some simple examples are given in the problems (2.4.18–19). In most of our discussions, we emphasize the Cramér–Rao bound.

We now digress briefly to develop a similar bound on the mean-square error when the parameter is random.

Lower Bound on the Minimum Mean-Square Error in Estimating a Random Parameter. In this section we prove the following theorem.

Theorem. Let a be a random variable and $\mathbf{r}$, the observation vector. The mean-square error of any estimate $\hat{a}(\mathbf{R})$ satisfies the inequality

$$E\{[\hat{a}(\mathbf{R}) - a]^2\} \geq \left(E\left\{\left[\frac{\partial \ln p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A}\right]^2\right\}\right)^{-1} = \left\{-E\left[\frac{\partial^2 \ln p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A^2}\right]\right\}^{-1}. \tag{217}$$

Observe that the probability density is a joint density and that the expectation is over both a and $\mathbf{r}$. The following conditions are assumed to exist:

1. $\dfrac{\partial p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A}$ is absolutely integrable with respect to $\mathbf{R}$ and A.
2. $\dfrac{\partial^2 p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A^2}$ is absolutely integrable with respect to $\mathbf{R}$ and A.
3. The conditional expectation of the error, given A, is

$$B(A) = \int_{-\infty}^{\infty} [\hat{a}(\mathbf{R}) - A]\, p_{\mathbf{r}|a}(\mathbf{R}|A)\, d\mathbf{R}. \tag{218}$$

We assume that

$$\lim_{A\to\infty} B(A)\, p_a(A) = 0, \tag{219}$$

$$\lim_{A\to-\infty} B(A)\, p_a(A) = 0. \tag{220}$$

The proof is a simple modification of the one on p. 66. Multiply both sides of (218) by $p_a(A)$ and then differentiate with respect to A:

$$\frac{d}{dA}[p_a(A)\, B(A)] = -\int_{-\infty}^{\infty} p_{\mathbf{r},a}(\mathbf{R}, A)\, d\mathbf{R} + \int_{-\infty}^{\infty} \frac{\partial p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A}[\hat{a}(\mathbf{R}) - A]\, d\mathbf{R}. \tag{221}$$

Now integrate with respect to A:

$$p_a(A)\, B(A)\Big|_{-\infty}^{+\infty} = -1 + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\partial p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A}[\hat{a}(\mathbf{R}) - A]\, dA\, d\mathbf{R}. \tag{222}$$

The assumption in Condition 3 makes the left-hand side zero. The remaining steps are identical. The result is

$$E\{[\hat{a}(\mathbf{R}) - a]^2\} \geq \left\{E\left[\left(\frac{\partial \ln p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A}\right)^2\right]\right\}^{-1} \tag{223}$$

or, equivalently,

$$E\{[\hat{a}(\mathbf{R}) - a]^2\} \geq \left\{-E\left[\frac{\partial^2 \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^2}\right] - E\left[\frac{\partial^2 \ln p_a(A)}{\partial A^2}\right]\right\}^{-1} \tag{224}$$

with equality if and only if

$$\frac{\partial \ln p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A} = k[\hat{a}(\mathbf{R}) - A], \tag{225}$$

for all $\mathbf{R}$ and all A. (In the nonrandom variable case we used the Schwarz inequality on an integral over $\mathbf{R}$ so that the constant $k(A)$ could be a function of A. Now the integration is over both $\mathbf{R}$ and A so that k cannot be a function of A.) Differentiating again gives an equivalent condition

$$\frac{\partial^2 \ln p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A^2} = -k. \tag{226}$$

Observe that (226) may be written in terms of the a posteriori density,

$$\frac{\partial^2 \ln p_{a|\mathbf{r}}(A|\mathbf{R})}{\partial A^2} = -k. \tag{227}$$

Integrating (227) twice and putting the result in the exponent, we have

$$p_{a|\mathbf{r}}(A|\mathbf{R}) = \exp(-kA^2 + C_1 A + C_2) \tag{228}$$

for all $\mathbf{R}$ and A; but (228) is simply a statement that the a posteriori probability density of a must be Gaussian for all $\mathbf{R}$ in order for an efficient estimate to exist. (Note that C_1 and C_2 are functions of $\mathbf{R}$).

Arguing as in (193)–(195), we see that if (226) is satisfied the MAP estimate will be efficient. Because the minimum MSE estimate cannot have a larger error, this tells us that $\hat{a}_{\text{ms}}(\mathbf{R}) = \hat{a}_{\text{map}}(\mathbf{R})$ whenever an efficient estimate exists. As a matter of technique, when an efficient estimate does exist, it is usually computationally easier to solve the MAP equation than it is to find the conditional mean. When an efficient estimate does not exist, we do not know how closely the mean-square error, using either $\hat{a}_{\text{ms}}(\mathbf{R})$ or $\hat{a}_{\text{map}}(\mathbf{R})$, approaches the lower bound. Asymptotic results similar to those for real variables may be derived.

2.4.3 Multiple Parameter Estimation

In many problems of interest we shall want to estimate more than one parameter. A familiar example is the radar problem in which we shall estimate the range and velocity of a target. Most of the ideas and techniques can be extended to this case in a straightforward manner. The model is shown in Fig. 2.23. If there are K parameters, $a_1, a_2, \ldots, a_K$, we describe them by a parameter vector $\mathbf{a}$ in a K-dimensional space. The other elements of the model are the same as before. We shall consider both the case in which $\mathbf{a}$ is a random parameter vector and that in which $\mathbf{a}$ is a real (or nonrandom) parameter vector. Three issues are of interest. In each the result is the vector analog to a result in the scalar case.

1. Estimation procedures.
2. Measures of error.
3. Bounds on performance.

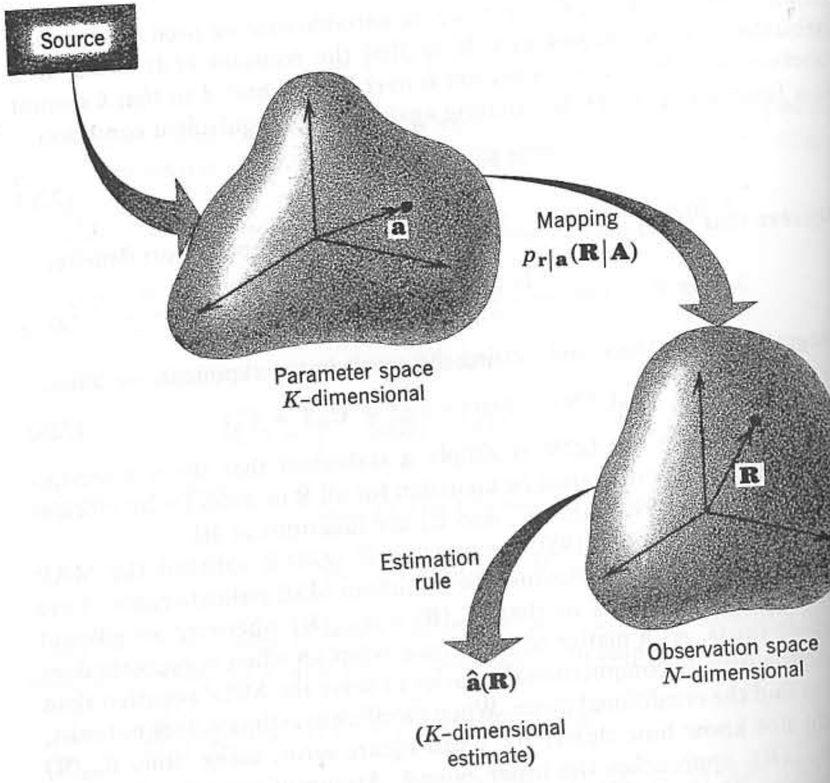


Fig. 2.23 Multiple parameter estimation model.

Estimation Procedure. For random variables we could consider the general case of Bayes estimation in which we minimize the risk for some arbitrary scalar cost function $C(\mathbf{a}, \hat{\mathbf{a}})$, but for our purposes it is adequate to consider only cost functions that depend on the error. We define the error vector as

$$\mathbf{a}_\epsilon(\mathbf{R}) = \begin{bmatrix} \hat{a}_1(\mathbf{R}) - a_1 \\ \hat{a}_2(\mathbf{R}) - a_2 \\ \vdots \\ \hat{a}_K(\mathbf{R}) - a_K \end{bmatrix} = \hat{\mathbf{a}}(\mathbf{R}) - \mathbf{a}. \tag{229}$$

For a mean-square error criterion, the cost function is simply

$$C(\mathbf{a}_\epsilon(\mathbf{R})) \triangleq \sum_{i=1}^{K} a_{\epsilon_i}{}^2(\mathbf{R}) = \mathbf{a}_\epsilon{}^T(\mathbf{R})\, \mathbf{a}_\epsilon(\mathbf{R}). \tag{230}$$

This is just the sum of the squares of the errors. The risk is

$$\mathcal{R}_{\text{ms}} = \iint_{-\infty}^{\infty} C(\mathbf{a}_\epsilon(\mathbf{R}))\, p_{\mathbf{r},\mathbf{a}}(\mathbf{R}, \mathbf{A})\, d\mathbf{R}\, d\mathbf{A} \tag{231}$$

or

$$\mathcal{R}_{\text{ms}} = \int_{-\infty}^{\infty} p_{\mathbf{r}}(\mathbf{R})\, d\mathbf{R} \int_{-\infty}^{\infty} \left[\sum_{i=1}^{K} (\hat{a}_i(\mathbf{R}) - A_i)^2\right] p_{\mathbf{a}|\mathbf{r}}(\mathbf{A}|\mathbf{R})\, d\mathbf{A}. \tag{232}$$

As before, we can minimize the inner integral for each $\mathbf{R}$. Because the terms in the sum are positive, we minimize them separately. This gives

$$\hat{a}_{\text{ms}_i}(\mathbf{R}) = \int_{-\infty}^{\infty} A_i p_{\mathbf{a}|\mathbf{r}}(\mathbf{A}|\mathbf{R})\, d\mathbf{A} \tag{233}$$

or

$$\hat{\mathbf{a}}_{\text{ms}}(\mathbf{R}) = \int_{-\infty}^{\infty} \mathbf{A} p_{\mathbf{a}|\mathbf{r}}(\mathbf{A}|\mathbf{R})\, d\mathbf{A}. \tag{234}$$

It is easy to show that mean-square estimation commutes over *linear transformations*. Thus, if

$$\mathbf{b} = \mathbf{D}\mathbf{a}, \tag{235}$$

where $\mathbf{D}$ is a $L \times K$ matrix, and we want to minimize

$$E[\mathbf{b}_\epsilon{}^T(\mathbf{R})\, \mathbf{b}_\epsilon(\mathbf{R})] = E\left[\sum_{i=1}^{L} b_{\epsilon_i}{}^2(\mathbf{R}),\right] \tag{236}$$

the result will be,

$$\hat{\mathbf{b}}_{\text{ms}}(\mathbf{R}) = \mathbf{D}\hat{\mathbf{a}}_{\text{ms}}(\mathbf{R}) \tag{237}$$

[see Problem 2.4.20 for the proof of (237)].

For MAP estimation we must find the value of **A** that maximizes $p_{\mathbf{a}|\mathbf{r}}(\mathbf{A}|\mathbf{R})$. If the maximum is interior and $\partial \ln p_{\mathbf{a}|\mathbf{r}}(\mathbf{A}|\mathbf{R})/\partial A_i$ exists at the maximum then a necessary condition is obtained from the MAP equations. By analogy with (137) we take the logarithm of $p_{\mathbf{a}|\mathbf{r}}(\mathbf{A}|\mathbf{R})$, differentiate with respect to each parameter A_i, $i = 1, 2, \ldots, K$, and set the result equal to zero. This gives a set of K simultaneous equations:

$$\left.\frac{\partial \ln p_{\mathbf{a}|\mathbf{r}}(\mathbf{A}|\mathbf{R})}{\partial A_i}\right|_{\mathbf{A}=\hat{\mathbf{a}}_{\text{map}}(\mathbf{R})} = 0, \qquad i = 1, 2, \ldots, K. \tag{238}$$

We can write (238) in a more compact manner by defining a partial derivative matrix operator

$$\nabla_{\mathbf{A}} \triangleq \begin{bmatrix} \dfrac{\partial}{\partial A_1} \\ \dfrac{\partial}{\partial A_2} \\ \vdots \\ \dfrac{\partial}{\partial A_K} \end{bmatrix}. \tag{239}$$

This operator can be applied only to $1 \times m$ matrices; for example,

$$\nabla_{\mathbf{A}}\,\mathbf{G} = \begin{bmatrix} \dfrac{\partial G_1}{\partial A_1} & \dfrac{\partial G_2}{\partial A_1} & \cdots & \dfrac{\partial G_m}{\partial A_1} \\ \vdots & & & \\ \dfrac{\partial G_1}{\partial A_K} & & & \dfrac{\partial G_m}{\partial A_K} \end{bmatrix}. \tag{240}$$

Several useful properties of $\nabla_{\mathbf{A}}$ are developed in Problems 2.4.27–28. In our case (238) becomes a single vector equation,

$$\nabla_{\mathbf{A}}[\ln p_{\mathbf{a}|\mathbf{r}}(\mathbf{A}|\mathbf{R})]|_{\mathbf{A}=\hat{\mathbf{a}}_{\text{map}}(\mathbf{R})} = \mathbf{0}. \tag{241}$$

Similarly, for ML estimates we must find the value of **A** that maximizes $p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})$. If the maximum is interior and $\partial \ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})|\partial A_i$ exists at the maximum then a necessary condition is obtained from the likelihood equations:

$$\nabla_{\mathbf{A}}[\ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})]|_{\mathbf{A}=\hat{\mathbf{a}}_{\text{ml}}(\mathbf{R})} = \mathbf{0}. \tag{242}$$

In both cases we must verify that we have the absolute maximum.

Measures of Error. For nonrandom variables the first measure of interest is the bias. Now the bias is a vector,

$$\mathbf{B}(\mathbf{A}) \triangleq E[\mathbf{a}_\epsilon(\mathbf{R})] = E[\hat{\mathbf{a}}(\mathbf{R})] - \mathbf{A}. \tag{243}$$

If each component of the bias vector is zero for every **A**, we say that the estimate is unbiased.

In the single parameter case a rough measure of the spread of the error was given by the variance of the estimate. In the special case in which $a_\epsilon(\mathbf{R})$ was Gaussian this provided a complete description:

$$p_{a_\epsilon}(A_\epsilon) = \frac{1}{\sqrt{2\pi}\,\sigma_{a_\epsilon}} \exp\left(-\frac{A_\epsilon^2}{2\sigma_{a_\epsilon}^2}\right). \tag{244}$$

For a vector variable the quantity analogous to the variance is the covariance matrix

$$E[(\mathbf{a}_\epsilon - \bar{\mathbf{a}}_\epsilon)(\mathbf{a}_\epsilon^T - \bar{\mathbf{a}}_\epsilon^T)] \triangleq \mathbf{\Lambda}_\epsilon, \tag{245}$$

where

$$\bar{\mathbf{a}}_\epsilon \triangleq E(\mathbf{a}_\epsilon) = \mathbf{B}(\mathbf{A}). \tag{246}$$

The best way to determine how the covariance matrix provides a measure of spread is to consider the special case in which the a_{ϵ_i} are jointly Gaussian. For algebraic simplicity we let $E(\mathbf{a}_\epsilon) = \mathbf{0}$. The joint probability density for a set of K jointly Gaussian variables is

$$p_{\mathbf{a}_\epsilon}(\mathbf{A}_\epsilon) = (|2\pi|^{K/2}|\mathbf{\Lambda}_\epsilon|^{1/2})^{-1} \exp\left(-\tfrac{1}{2}\mathbf{A}_\epsilon^T\mathbf{\Lambda}_\epsilon^{-1}\mathbf{A}_\epsilon\right) \tag{247}$$

(e.g., p. 151 in Davenport and Root [1]).

The probability density for $K = 2$ is shown in Fig. 2.24*a*. In Figs. 2.24*b,c* we have shown the equal-probability contours of two typical densities.From (247) we observe that the equal-height contours are defined by the relation,

$$\mathbf{A}_\epsilon^T\mathbf{\Lambda}_\epsilon^{-1}\mathbf{A}_\epsilon = C^2, \tag{248}$$

which is the equation for an ellipse when $K = 2$. The ellipses move out monotonically with increasing C. They also have the interesting property that the probability of being inside the ellipse is only a function of C^2.

Property. For $K = 2$, the probability that the error vector lies inside an ellipse whose equation is

$$\mathbf{A}_\epsilon^T\mathbf{\Lambda}_\epsilon^{-1}\mathbf{A}_\epsilon = C^2, \tag{249}$$

is

$$P = 1 - \exp\left(-\frac{C^2}{2}\right). \tag{250}$$

Proof. The area inside the ellipse defined by (249) is

$$\mathcal{A} = |\mathbf{\Lambda}_\epsilon|^{1/2}\pi C^2. \tag{251}$$

The differential area between ellipses corresponding to C and $C + dC$ respectively is

$$d\mathcal{A} = |\mathbf{\Lambda}_\epsilon|^{1/2} 2\pi C\, dC. \tag{252}$$

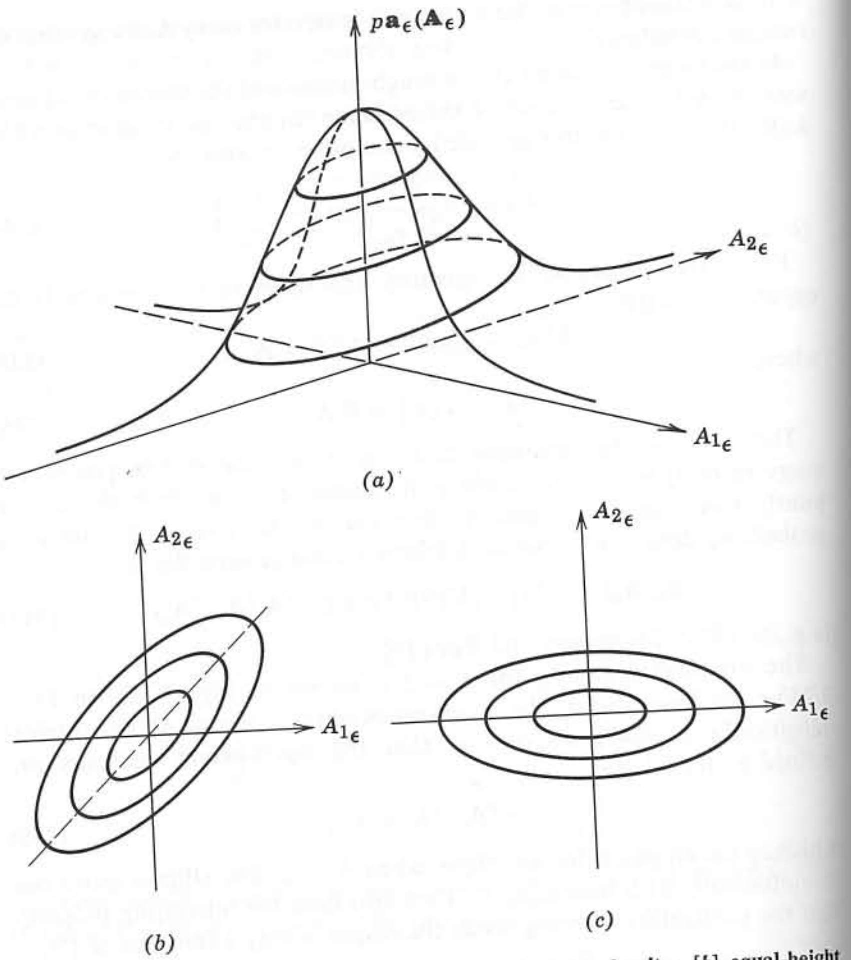


Fig. 2.24 Gaussian densities: [*a*] two-dimensional Gaussian density; [*b*] equal-height contours, correlated variables; [*c*] equal-height contours, uncorrelated variables.

The height of the probability density in this differential area is

$$(2\pi|\mathbf{\Lambda}_\epsilon|^{1/2})^{-1}\exp\left(-\frac{C^2}{2}\right). \tag{253}$$

We can compute the probability of a point lying outside the ellipse by multiplying (252) by (253) and integrating from C to ∞.

$$1 - P = \int_C^\infty X\exp\left(-\frac{X^2}{2}\right)dX = \exp\left(-\frac{C^2}{2}\right), \tag{254}$$

which is the desired result.

For this reason the ellipses described by (248) are referred to as *concentration ellipses* because they provide a measure of the concentration of the density.

A similar result holds for arbitrary K. Now, (248) describes an *ellipsoid*. Here the differential volume† in K-dimensional space is

$$dv = |\mathbf{\Lambda}_\epsilon|^{1/2}\frac{\pi^{K/2}}{\Gamma(K/2+1)}KC^{K-1}\,dC. \tag{255}$$

The value of the probability density on the ellipsoid is

$$[(2\pi)^{K/2}|\mathbf{\Lambda}_\epsilon|^{1/2}]^{-1}\exp\left(-\frac{C^2}{2}\right). \tag{256}$$

Therefore

$$1 - P = \frac{K}{(2)^{K/2}\Gamma(K/2+1)}\int_C^\infty X^{K-1}e^{-X^2/2}\,dX, \tag{257}$$

which is the desired result. We refer to these ellipsoids as *concentration ellipsoids*.

When the probability density of the error is *not* Gaussian, the concentration ellipsoid no longer specifies a unique probability. This is directly analogous to the one-dimensional case in which the variance of a non-Gaussian zero-mean random variable does not determine the probability density. We can still interpret the concentration ellipsoid as a rough measure of the spread of the errors. When the concentration ellipsoids of a given density lie wholly outside the concentration ellipsoids of a second density, we say that the second density is more concentrated than the first. With this motivation, we derive some properties and bounds pertaining to concentration ellipsoids.

Bounds on Estimation Errors: Nonrandom Variables. In this section we derive two bounds. The first relates to the variance of an individual error; the second relates to the concentration ellipsoid.

Property 1. Consider *any* unbiased estimate of A_i. Then

$$\sigma_{\epsilon_i}{}^2 \triangleq \operatorname{Var}[\hat{a}_i(\mathbf{R}) - A_i] \geq J^{ii}, \tag{258}$$

where J^{ii} is the iith element in the $K \times K$ square matrix $\mathbf{J}^{-1}$. The elements in $\mathbf{J}$ are

$$\begin{aligned} J_{ij} &\triangleq E\left[\frac{\partial \ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})}{\partial A_i}\cdot\frac{\partial \ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})}{\partial A_j}\right] \\ &= -E\left[\frac{\partial^2 \ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})}{\partial A_i\,\partial A_j}\right] \end{aligned} \tag{259}$$

† e.g., Cramér [9], p. 120, or Sommerfeld [32].

or

$$\begin{aligned}\mathbf{J} &\triangleq E(\{\nabla_{\mathbf{A}}[\ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})]\}\{\nabla_{\mathbf{A}}[\ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})]\}^T) \\ &= -E[\nabla_{\mathbf{A}}(\{\nabla_{\mathbf{A}}[\ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})]\}^T)].\end{aligned} \tag{260}$$

The **J** matrix is commonly called *Fisher's information matrix*. The equality in (258) holds if and only if

$$\hat{a}_i(\mathbf{R}) - A_i = \sum_{j=1}^{K} k_{ij}(\mathbf{A}) \frac{\partial \ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})}{\partial A_j} \tag{261}$$

for all values of A_i and **R**.

In other words, the estimation error can be expressed as the weighted sum of the partial derivatives of $\ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})$ with respect to the various parameters.

Proof. Because $\hat{a}_i(\mathbf{R})$ is unbiased,

$$\int_{-\infty}^{\infty} [\hat{a}_i(\mathbf{R}) - A_i]\, p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})\, d\mathbf{R} = 0 \tag{262}$$

or

$$\int_{-\infty}^{\infty} \hat{a}_i(\mathbf{R})\, p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})\, d\mathbf{R} = A_i. \tag{263}$$

Differentiating both sides with respect to A_j, we have

$$\begin{aligned}&\int_{-\infty}^{\infty} \hat{a}_i(\mathbf{R}) \frac{\partial p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})}{\partial A_j}\, d\mathbf{R} \\ &\qquad = \int_{-\infty}^{\infty} \hat{a}_i(\mathbf{R}) \frac{\partial \ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})}{\partial A_j}\, p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})\, d\mathbf{R} = \delta_{ij}.\end{aligned} \tag{264}$$

We shall prove the result for $i = 1$. We define a $K + 1$ vector

$$\mathbf{x} = \begin{bmatrix} \hat{a}_1(\mathbf{R}) - A_1 \\ \dfrac{\partial \ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})}{\partial A_1} \\ \vdots \\ \dfrac{\partial \ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})}{\partial A_K} \end{bmatrix}. \tag{265}$$

The covariance matrix is

$$E[\mathbf{x}\mathbf{x}^T] = \begin{bmatrix} \sigma_{\epsilon_1}^{2} & 1 & 0 & 0 & 0 \\ 1 & J_{11} & J_{12} & \cdots & J_{1K} \\ 0 & \vdots & & \ddots & \vdots \\ 0 & J_{K1} & & & J_{KK} \end{bmatrix}. \tag{266}$$

[The ones and zeroes in the matrix follow from (264).] Because it is a covariance matrix, it is nonnegative definite, which implies that the determinant of the entire matrix is greater than or equal to zero. (This condition is only necessary, not sufficient, for the matrix to be nonnegative definite.)

Evaluating the determinant using a cofactor expansion, we have

$$\sigma_{\epsilon_1}^{2}|\mathbf{J}| - \text{cofactor } J_{11} \geq 0. \tag{267}$$

If we assume that **J** is nonsingular, then

$$\sigma_{\epsilon_1}^{2} \geq \frac{\text{cofactor } J_{11}}{|\mathbf{J}|} = J^{11}, \tag{268}$$

which is the desired result. The modifications for the case when **J** is singular follow easily for any specific problem.

In order for the determinant to equal zero, the term $\hat{A}_1(\mathbf{R}) - A_1$ must be expressible as a linear combination of the other terms. This is the condition described by (261). The second line of (259) follows from the first line in a manner exactly analogous to the proof in (189)–(192). The proof for $i \neq 1$ is an obvious modification.

Property 2. Consider *any* unbiased estimate of **A**. The concentration ellipse

$$\mathbf{A}_\epsilon{}^T \mathbf{\Lambda}_\epsilon{}^{-1} \mathbf{A}_\epsilon = C^2 \tag{269}$$

lies either outside or on the bound ellipse defined by

$$\mathbf{A}_\epsilon{}^T \mathbf{J} \mathbf{A}_\epsilon = C^2. \tag{270}$$

Proof. We shall go through the details for $K = 2$. By analogy with the preceding proof, we construct the covariance matrix of the vector.

$$\mathbf{x} = \begin{bmatrix} \hat{a}_1(\mathbf{R}) - A_1 \\ \hat{a}_2(\mathbf{R}) - A_2 \\ \dfrac{\partial \ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})}{\partial A_1} \\ \dfrac{\partial \ln p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})}{\partial A_2} \end{bmatrix}. \tag{271}$$

Then

$$E[\mathbf{x}\mathbf{x}^T] = \left[\begin{array}{cc|cc} \sigma_{1_\epsilon}^{2} & \rho\sigma_{1_\epsilon}\sigma_{2_\epsilon} & 1 & 0 \\ \rho\sigma_{1_\epsilon}\sigma_{2_\epsilon} & \sigma_{2_\epsilon}^{2} & 0 & 1 \\ \hline 1 & 0 & J_{11} & J_{12} \\ 0 & 1 & J_{21} & J_{22} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{\Lambda}_\epsilon & \mathbf{I} \\ \hline \mathbf{I} & \mathbf{J} \end{array}\right]. \tag{272}$$

The second equality defines a partition of the 4 × 4 matrix into four 2 × 2 matrices. Because it is a covariance matrix, it is nonnegative definite. Using a formula for the determinant of a partitioned matrix,† we have

$$|\mathbf{\Lambda}_\epsilon \mathbf{J} - \mathbf{I}| \geq 0 \tag{273}$$

or, assuming that $\mathbf{\Lambda}_\epsilon$ is nonsingular and applying the product rule for determinants,

$$|\mathbf{\Lambda}_\epsilon|\ |\mathbf{J} - \mathbf{\Lambda}_\epsilon^{-1}| \geq 0. \tag{274}$$

This implies

$$|\mathbf{J} - \mathbf{\Lambda}_\epsilon^{-1}| \geq 0. \tag{275}$$

Now consider the two ellipses. The intercept on the A_{ϵ_1} axis is

$$A_{1_\epsilon}^{2}\Big|_{A_{2\epsilon}=0} = C^2 \frac{|\mathbf{\Lambda}_\epsilon|}{\sigma_2^{2}} \tag{276}$$

for the actual concentration ellipse and

$$A_{1_\epsilon}^{2}\Big|_{A_{2\epsilon}=0} = C^2 \frac{1}{J_{11}} \tag{277}$$

for the bound ellipse.

We want to show that the actual intercept is greater than or equal to the bound intercept. This requires

$$J_{11}|\mathbf{\Lambda}_\epsilon| \geq \sigma_2^{2}. \tag{278}$$

This inequality follows because the determinant of the 3 × 3 matrix in the upper left corner of (272) is greater than or equal to zero. (Otherwise the entire matrix is not nonnegative definite, e.g. [16] or [18].) Similarly, the actual intercept on the A_{2_ϵ} axis is greater than or equal to the bound intercept. Therefore the actual ellipse is either always outside (or on) the bound ellipse *or* the two ellipses intersect.

If they intersect, we see from (269) and (270) that there must be a solution, $\mathbf{A}_\epsilon$, to the equation

$$\mathbf{A}_\epsilon^T \mathbf{\Lambda}_\epsilon^{-1} \mathbf{A}_\epsilon = \mathbf{A}_\epsilon^T \mathbf{J} \mathbf{A}_\epsilon \tag{279}$$

or

$$\mathbf{A}_\epsilon^T[\mathbf{J} - \mathbf{\Lambda}_\epsilon^{-1}]\mathbf{A}_\epsilon \triangleq \mathbf{A}_\epsilon^T \mathbf{D} \mathbf{A}_\epsilon = 0. \tag{280}$$

In scalar notation

$$A_{1_\epsilon}^{2} D_{11} + 2A_{1_\epsilon}A_{2_\epsilon}D_{12} + A_{2_\epsilon}^{2} D_{22} = 0 \tag{281}$$

or, equivalently,

$$\left(\frac{A_{1_\epsilon}}{A_{2_\epsilon}}\right)^2 D_{11} + 2\left(\frac{A_{1_\epsilon}}{A_{2_\epsilon}}\right) D_{12} + D_{22} = 0. \tag{282}$$

† Bellman [16], p. 83.

Solving for $A_{1_\epsilon}/A_{2_\epsilon}$, we would obtain real roots only if the discriminant were greater than or equal to zero. This requires

$$|\mathbf{J} - \mathbf{\Lambda}_\epsilon^{-1}| \leq 0. \tag{283}$$

The inequality is a contradiction of (275). One possibility is $|\mathbf{J} - \mathbf{\Lambda}_\epsilon^{-1}| = 0$, but this is true only when the ellipses coincide. In this case all the estimates are efficient.

For arbitrary K we can show that $\mathbf{J} - \mathbf{\Lambda}_\epsilon^{-1}$ is nonnegative definite. The implications with respect to the concentration ellipsoids are the same as for $K = 2$.

Frequently we want to estimate functions of the K basic parameters rather than the parameters themselves. We denote the desired estimates as

$$\begin{aligned} d_1 &= g_{d_1}(\mathbf{A}), \\ d_2 &= g_{d_2}(\mathbf{A}), \\ &\vdots \\ d_M &= g_{d_M}(\mathbf{A}). \end{aligned} \tag{284}$$

or

$$\mathbf{d} = \mathbf{g}_d(\mathbf{A})$$

The number of estimates M is not related to K in general. The functions may be nonlinear. The estimation error is

$$\hat{d}_i - g_i(\mathbf{A}) \triangleq d_{\epsilon_i}. \tag{285}$$

If we assume that the estimates are unbiased and denote the error covariance matrix as $\mathbf{\Lambda}_\epsilon$, then by using methods identical to those above we can prove the following properties.

Property 3. The matrix

$$\mathbf{\Lambda}_\epsilon - \{\nabla_{\mathbf{A}}[\mathbf{g}_d^T(\mathbf{A})]\}^T \mathbf{J}^{-1}\{\nabla_{\mathbf{A}}[\mathbf{g}_d^T(\mathbf{A})]\} \tag{286}$$

is nonnegative definite.

This implies the following property (just multiply the second matrix out and recall that all diagonal elements of nonnegative definite matrix are nonnegative):

Property 4.

$$\text{Var}\,(d_{\epsilon_l}) \geq \sum_{i}^{K} \sum_{j}^{K} \frac{\partial g_{d_l}(\mathbf{A})}{\partial A_i} J^{ij} \frac{\partial g_{d_l}(\mathbf{A})}{\partial A_j}. \tag{287}$$

For the special case in which the desired functions are linear, the result in (287) can be written in a simpler form.

Property 5. Assume that

$$\mathbf{g}_d(\mathbf{A}) \triangleq \mathbf{G}_d\mathbf{A}, \tag{288}$$

where $\mathbf{G}_d$ is an $M \times K$ matrix. If the estimates are unbiased, then

$$\mathbf{\Lambda}_\epsilon - \mathbf{G}_d\mathbf{J}^{-1}\mathbf{G}_d{}^T$$

is nonnegative definite.

Property 6. Efficiency commutes with linear transformations but does not commute with nonlinear transformations. In other words, if $\hat{\mathbf{A}}$ is efficient, then $\mathbf{g}_d(\mathbf{A})$ will be efficient if and only if $\mathbf{g}_d(\mathbf{A})$ is a linear transformation.

Bounds on Estimation Errors: Random Parameters. Just as in the single parameter case, the bound for random parameters is derived by a straightforward modification of the derivation for nonrandom parameters. The information matrix now consists of two parts:

$$\mathbf{J}_T \triangleq \mathbf{J}_D + \mathbf{J}_P. \tag{289}$$

The matrix $\mathbf{J}_D$ is the information matrix defined in (260) and represents information obtained from the *data*. The matrix $\mathbf{J}_P$ represents the a priori information. The elements are

$$J_{P_{ij}} \triangleq E\left[\frac{\partial \ln p_{\mathbf{a}}(\mathbf{A})}{\partial A_i}\,\frac{\partial \ln p_{\mathbf{a}}(\mathbf{A})}{\partial A_j}\right] \tag{290}$$

$$= -E\left[\frac{\partial^2 \ln p_{\mathbf{a}}(\mathbf{A})}{\partial A_i\, \partial A_j}\right].$$

The *correlation matrix* of the errors is

$$\mathbf{R}_\epsilon \triangleq E(\mathbf{a}_\epsilon\mathbf{a}_\epsilon{}^T). \tag{291}$$

The diagonal elements represent the mean-square errors and the off-diagonal elements are the cross correlations. Three properties follow easily:

Property No. 1.

$$E[a_{\epsilon_i}{}^2] \geq J_T{}^{ii}. \tag{292}$$

In other words, the diagonal elements in the inverse of the total information matrix are lower bounds on the corresponding mean-square errors.

Property No. 2. The matrix

$$\mathbf{J}_T - \mathbf{R}_\epsilon{}^{-1}$$

is nonnegative definite. This has the same physical interpretation as in the nonrandom parameter problem.

Property No. 3. If $\mathbf{J}_T = \mathbf{R}_\epsilon{}^{-1}$, all of the estimates are efficient. A necessary and sufficient condition for this to be true is that $p_{\mathbf{a}|\mathbf{r}}(\mathbf{A}|\mathbf{R})$ be Gaussian for all $\mathbf{R}$. This will be true iff $\mathbf{J}$ is constant. [Modify (261), (228)].

A special case of interest occurs when the a priori density is a Kth-order Gaussian density. Then

$$\mathbf{J}_P = \mathbf{\Lambda}_{\mathbf{a}}{}^{-1}, \tag{293}$$

where $\mathbf{\Lambda}_{\mathbf{a}}$ is the covariance matrix of the random parameters.

An even simpler case arises when the variables are independent Gaussian variables. Then

$$J_{P_{ij}} = \frac{1}{\sigma_{a_i}{}^2}\,\delta_{ij}, \tag{294}$$

Under these conditions only the diagonal terms of $\mathbf{J}_T$ are affected by the a priori information.

Results similar to Properties 3 to 6 for nonrandom parameters can be derived for the random parameter case.

2.4.4 Summary of Estimation Theory

In this section we developed the estimation theory results that we shall need for the problems of interest. We began our discussion with Bayes estimation of random parameters. The basic quantities needed in the model were the a priori density $p_a(A)$, the probabilistic mapping to the observation space $p_{\mathbf{r}|a}(\mathbf{R}|A)$, and a cost function $C(A_\epsilon)$. These quantities enabled us to find the risk. The estimate which minimized the risk was called a Bayes estimate and the resulting risk, the Bayes risk. Two types of Bayes estimate, the MMSE estimate (which was the mean of the a posteriori density) and the MAP estimate (the mode of the a posteriori density), were emphasized. In Properties 1 and 2 (pp. 60–61) we saw that the conditional mean was the Bayes estimate for a large class of cost functions when certain conditions on the cost function and a posteriori density were satisfied.

Turning to nonrandom parameter estimation, we introduced the idea of bias and variance as two separate error measures. The Cramér-Rao inequality provided a bound on the variance of any unbiased estimate. Whenever an efficient estimate existed, the maximum likelihood estimation procedure gave this estimate. This property of the ML estimate, coupled with its asymptotic properties, is the basis for our emphasis on ML estimates.

The extension to multiple parameter estimation involved no new concepts. Most of the properties were just multidimensional extensions of the corresponding scalar result.

It is important to emphasize the close relationship between detection and estimation theory. Both theories are based on a likelihood function or likelihood ratio, which, in turn, is derived from the probabilistic transition mechanism. As we proceed to more difficult problems, we shall find that a large part of the work is the manipulation of this transition mechanism. In many cases the mechanism will not depend on whether the problem is one of detection or estimation. Thus the difficult part of the problem will be applicable to either problem. This close relationship will become even more obvious as we proceed. We now return to the detection theory problem and consider a more general model.

2.5 COMPOSITE HYPOTHESES

In Sections 2.2 and 2.3 we confined our discussion to the decision problem in which the hypotheses were simple. We now extend our discussion to the case in which the hypotheses are composite. The term composite is most easily explained by a simple example.

Example 1. Under hypothesis 0 the observed variable r is Gaussian with zero mean and variance σ^2. Under hypothesis 1 the observed variable r is Gaussian with mean m and variance σ^2. The value of m can be anywhere in the interval $[M_0, M_1]$. Thus

$$H_0: p_{r|H_0}(R|H_0) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R^2}{2\sigma^2}\right), \tag{295}$$

$$H_1: p_{r|H_1}(R|H_1) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(R-M)^2}{2\sigma^2}\right), \qquad M_0 \le M \le M_1.$$

We refer to H_1 as a composite hypothesis because the parameter value M, which characterizes the hypothesis, ranges over a set of values. A model of this decision problem is shown in Fig. 2.25*a*. The output of the source is a parameter value M, which we view as a point in a parameter space χ. We then define the hypotheses as subspaces of χ. In this case H_0 corresponds to the point $M = 0$ and H_1 corresponds to the interval $[M_0, M_1.]$ We assume that the probability density governing the mapping from the parameter space to the observation space, $p_{r|m}(R|M)$, is known for all values of M in χ.

The final component is a decision rule that divides the observation space into two parts which correspond to the two possible decisions. It is important to observe that we are interested *solely* in making a decision and that the actual value of M is not of interest to us. For this reason the parameter M is frequently referred to as an "unwanted" parameter.

The extension of these ideas to the general composite hypothesis-testing problem is straightforward. The model is shown in Fig. 2.25*b*. The output of the source is a set of parameters. We view it as a point in a parameter space χ and denote it by the vector $\boldsymbol{\theta}$. The hypotheses are subspaces of χ. (In Fig. 2.25*b* we have indicated nonoverlapping spaces for convenience.) The probability density governing the mapping from the parameter space to the observation space is denoted by $p_{\mathbf{r}|\boldsymbol{\theta}}(\mathbf{R}|\boldsymbol{\theta})$ and is assumed to be known for all values of $\boldsymbol{\theta}$ in χ. Once again, the final component is a decision rule.

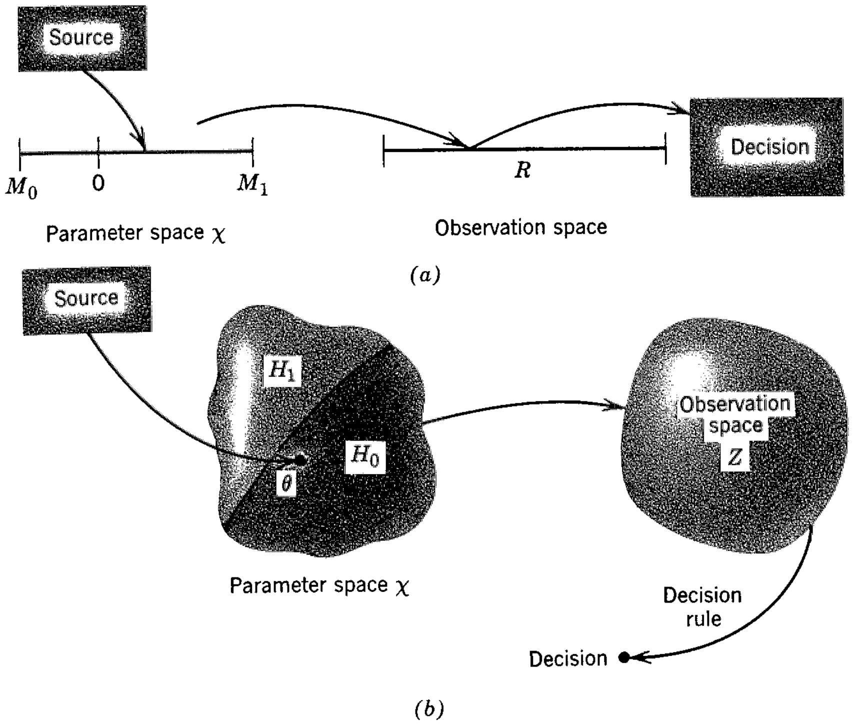


Fig. 2.25 ***a.*** **Composite hypothesis testing problem for single-parameter example.** ***b.*** **Composite hypothesis testing problem.**

To complete the formulation, we must characterize the parameter $\boldsymbol{\theta}$. Just as in the parameter estimation case the parameter $\boldsymbol{\theta}$ may be a nonrandom or random variable. If $\boldsymbol{\theta}$ is a random variable with a known probability density, the procedure is straightforward. Denoting the probability density of $\boldsymbol{\theta}$ on the two hypotheses as $p_{\boldsymbol{\theta}|H_0}(\boldsymbol{\theta}|H_0)$ and $p_{\boldsymbol{\theta}|H_1}(\boldsymbol{\theta}|H_1)$, the likelihood ratio is

$$\Lambda(\mathbf{R}) \triangleq \frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)} = \frac{\displaystyle\int_\chi p_{\mathbf{r}|\boldsymbol{\theta}}(\mathbf{R}|\boldsymbol{\theta})\, p_{\boldsymbol{\theta}|H_1}(\boldsymbol{\theta}|H_1)\, d\boldsymbol{\theta}}{\displaystyle\int_\chi p_{\mathbf{r}|\boldsymbol{\theta}}(\mathbf{R}|\boldsymbol{\theta})\, p_{\boldsymbol{\theta}|H_0}(\boldsymbol{\theta}|H_0)\, d\boldsymbol{\theta}}. \tag{296}$$

The reason for this simplicity is that the known probability density on $\boldsymbol{\theta}$ enables us to reduce the problem to a simple hypothesis-testing problem by integrating over $\boldsymbol{\theta}$. We can illustrate this procedure for the model in Example 1.

Example 1 (*continued.*) We assume that the probability density governing m on H_1 is

$$p_{m|H_1}(M|H_1) = \frac{1}{\sqrt{2\pi}\,\sigma_m} \exp\left(-\frac{M^2}{2\sigma_m{}^2}\right), \qquad -\infty < M < \infty, \tag{297}$$

Then (296) becomes

$$\Lambda(R) = \frac{\displaystyle\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(R-M)^2}{2\sigma^2}\right) \cdot \frac{1}{\sqrt{2\pi}\,\sigma_m} \exp\left(-\frac{M^2}{2\sigma_m{}^2}\right) dM}{\displaystyle\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R^2}{2\sigma^2}\right)} \mathop{\gtrless}_{H_0}^{H_1} \eta. \tag{298}$$

Integrating and taking the logarithm of both sides, we obtain

$$R^2 \mathop{\gtrless}_{H_0}^{H_1} \frac{2\sigma^2(\sigma^2 + \sigma_m{}^2)}{\sigma_m{}^2} \left[\ln \eta + \frac{1}{2} \ln \left(1 + \frac{\sigma_m{}^2}{\sigma^2}\right)\right]. \tag{299}$$

This result is equivalent to Example 2 on p. 29 because the density used in (297) makes the two problems identical.

As we expected, the test uses only the magnitude of R because the mean m has a symmetric probability density.

For the general case given in (296) the actual calculation may be more involved, but the desired procedure is well defined.

When $\boldsymbol{\theta}$ is a random variable with an unknown density, the best test procedure is not clearly specified. One possible approach is a minimax test over the unknown density. An alternate approach is to try several densities based on any partial knowledge of $\boldsymbol{\theta}$ that is available. In many cases the test structure will be insensitive to the detailed behavior of the probability density.

The second case of interest is the case in which $\boldsymbol{\theta}$ is a nonrandom variable. Here, just as in the problem of estimating nonrandom variables, we shall try a procedure and investigate the results. A first observation is that, because $\boldsymbol{\theta}$ has no probability density over which to average, a Bayes test is not meaningful. Thus we can devote our time to Neyman-Pearson tests.

We begin our discussion by examining what we call a *perfect measurement* bound on the test performance. We illustrate this idea for the problem in Example 1.

Example 2. In this case $\theta = M$.

From (295)

$$H_1 : p_{r|m}(R|M) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(R-M)^2}{2\sigma^2}\right), \qquad (M_0 \le M \le M_1), \tag{300}$$

and

$$H_0 : p_{r|m}(R|M) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R^2}{2\sigma^2}\right).$$

where M is an unknown nonrandom parameter.

It is clear that whatever test we design can never be better than a hypothetical test in which the receiver first measures M perfectly (or, alternately, it is told M) and then designs the optimum likelihood ratio test. Thus we can bound the ROC of any test by the ROC of this fictitious perfect measurement test. For this example we could use the ROC's in Fig. 2.9*a* by letting $d^2 = M^2/\sigma^2$. Because we are interested in the behavior versus M, the format in Fig. 2.9*b* is more useful. This is shown in Fig. 2.26. Such a curve is called a *power function*. It is simply a plot of P_D for all values of M (more generally $\boldsymbol{\theta}$) for various values of P_F. Because $H_0 = H_1$ for $M = 0$, $P_D = P_F$. The curves in Fig. 2.26 represent a bound on how well any test could do. We now want to see how close the actual test performance comes to this bound.

The best performance we could achieve would be obtained if an actual test's curves equaled the bound for all $M \in \chi$. We call such tests *uniformly most powerful* (UMP). In other words, for a given P_F a UMP test has a P_D greater than or equal to any other test for *all* $M \in \chi$. The conditions for a UMP test to exist can be seen in Fig. 2.27.

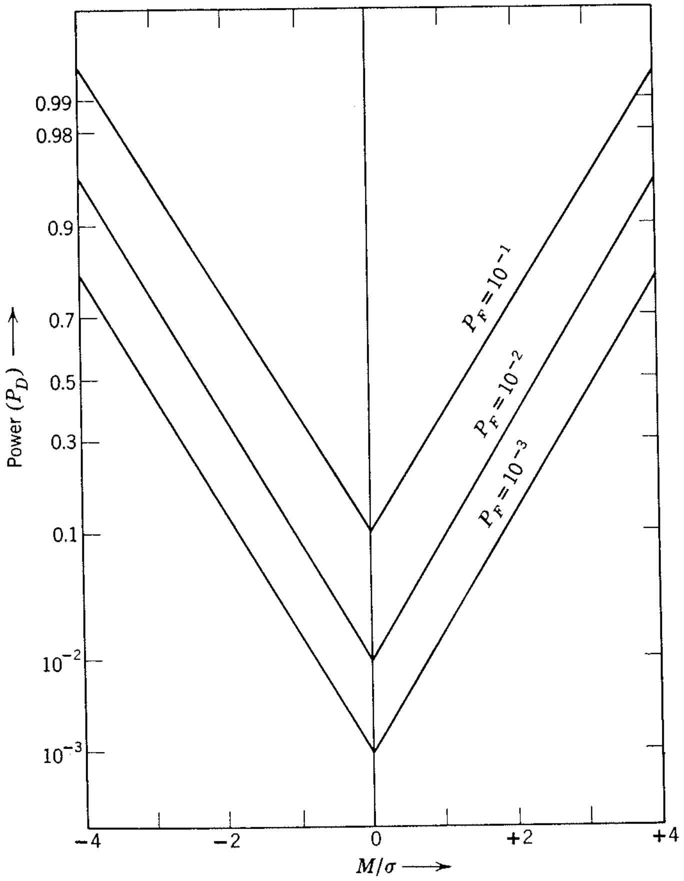


Fig. 2.26 Power function for perfect measurement test.

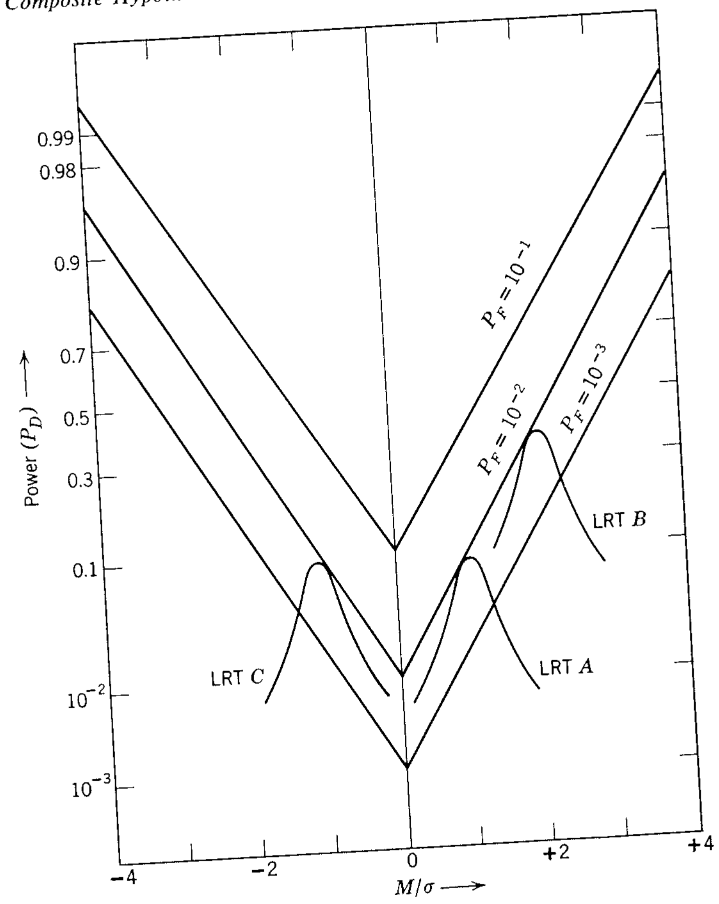


Fig. 2.27 Power functions for various likelihood ratio tests.

We first construct the perfect measurement bound. We next consider other possible tests and their performances. Test A is an ordinary likelihood ratio test designed under the assumption that $M = 1$. The first observation is that the power of this test equals the bound at $M = 1$, which follows from the manner in which we constructed the bound. For other values of M the power of test A may or may not equal the bound. Similarly, test B is a likelihood ratio test designed under the assumption that $M = 2$, and test C is a likelihood ratio test designed under the assumption that $M = -1$. In each case their power equals the bound at their design points. (The power functions in Fig. 2.27 are drawn to emphasize this and are not quantitatively correct away from the design point. The quantitatively correct curves are shown in Fig. 2.29.) They may also equal the bound at other points. The conditions for a UMP test are now obvious. We must be able to design a complete likelihood ratio test (including the threshold) for every $M \in \chi$ without knowing M.

The analogous result for the general case follows easily.

It is clear that in general the bound can be reached for any particular $\boldsymbol{\theta}$ simply by designing an ordinary LRT for that particular $\boldsymbol{\theta}$. Now a UMP test must be as good as any other test for every $\boldsymbol{\theta}$. This gives us a necessary and sufficient condition for its existence.

Property. A UMP test exists if and only if the likelihood ratio test for every $\boldsymbol{\theta} \in \chi$ can be completely defined (including threshold) without knowledge of $\boldsymbol{\theta}$.

The "if" part of the property is obvious. The "only if" follows directly from our discussion in the preceding paragraph. If there exists some $\boldsymbol{\theta} \in \chi$ for which we cannot find the LRT without knowing $\boldsymbol{\theta}$, we should have to use some other test, because we do not know $\boldsymbol{\theta}$. This test will necessarily be inferior for that particular $\boldsymbol{\theta}$ to a LRT test designed for that particular θ and therefore is not *uniformly* most powerful.

Returning to our example and using the results in Fig. 2.8, we know that the likelihood ratio test is

$$R \underset{H_0}{\overset{H_1}{\gtrless}} \gamma^+, \tag{301}$$

and

$$P_F = \int_{\gamma^+}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R^2}{2\sigma^2}\right) dR, \qquad \text{if } M > 0. \tag{302}$$

(The superscript + emphasizes the test assumes $M > 0$. The value of γ^+ may be negative.) This is shown in Fig. 2.28*a*.

Similarly, for the case in which $M < 0$ the likelihood ratio test is

$$R \underset{H_1}{\overset{H_0}{\gtrless}} \gamma^-, \tag{303}$$

where

$$P_F = \int_{-\infty}^{\gamma^-} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R^2}{2\sigma^2}\right) dR, \qquad M < 0. \tag{304}$$

This is shown in Fig. 2.28*b*. We see that the threshold is just the negative of the threshold for $M > 0$. This reversal is done to get the largest portion of $p_{r|H_1}(R|H_1)$ inside the H_1 region (and therefore maximize P_D).

Thus, with respect to Example 1, we draw the following conclusions:

1. If M can take on *only* nonnegative values (i.e., $M_0 \geq 0$), a UMP test exists [use (301)].
2. If M can take on *only* nonpositive values (i.e., $M_1 \leq 0$), a UMP test exists [use (303)].
3. If M can take on both negative and positive values (i.e., $M_0 < 0$ and $M_1 > 0$), then a UMP test does not exist. In Fig. 2.29 we show the power function for a likelihood ratio test designed under the assumption that M was positive. For negative values of M, P_D is less than P_F because the threshold is on the wrong side.

Whenever a UMP test exists, we use it, and the test works as well as if we knew $\boldsymbol{\theta}$. A more difficult problem is presented when a UMP test does

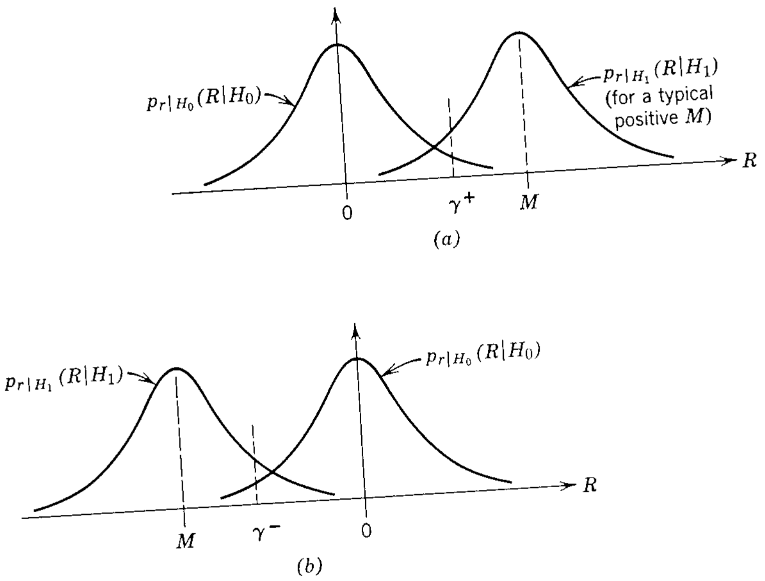


Fig. 2.28 Effect of sign of M: [a] threshold for positive M; [b] threshold for negative M.

not exist. The next step is to discuss other possible tests for the cases in which a UMP test does not exist. We confine our discussion to one possible test procedure. Others are contained in various statistics texts (e.g., Lehmann [17]) but seem to be less appropriate for the physical problems of interest in the sequel.

The perfect measurement bound suggests that a logical procedure is to estimate $\boldsymbol{\theta}$ assuming H_1 is true, then estimate $\boldsymbol{\theta}$ assuming H_0 is true, and use these estimates in a likelihood ratio test as if they were correct. If the maximum likelihood estimates discussed on p. 65 are used, the result is called a *generalized likelihood ratio test*. Specifically,

$$\Lambda_g(\mathbf{R}) = \frac{\max\limits_{\boldsymbol{\theta}_1} p_{\mathbf{r}|\boldsymbol{\theta}_1}(\mathbf{R}|\boldsymbol{\theta}_1)}{\max\limits_{\boldsymbol{\theta}_0} p_{\mathbf{r}|\boldsymbol{\theta}_0}(\mathbf{R}|\boldsymbol{\theta}_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma, \tag{305}$$

where $\boldsymbol{\theta}_1$ ranges over all $\boldsymbol{\theta}$ in H_1 and $\boldsymbol{\theta}_0$ ranges over all $\boldsymbol{\theta}$ in H_0. In other words, we make a ML estimate of $\boldsymbol{\theta}_1$, assuming that H_1 is true. We then evaluate $p_{\mathbf{r}|\boldsymbol{\theta}_1}(\mathbf{R}|\boldsymbol{\theta}_1)$ for $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1$ and use this value in the numerator. A similar procedure gives the denominator.

A simple example of a generalized LRT is obtained by using a slightly modified version of Example 1.

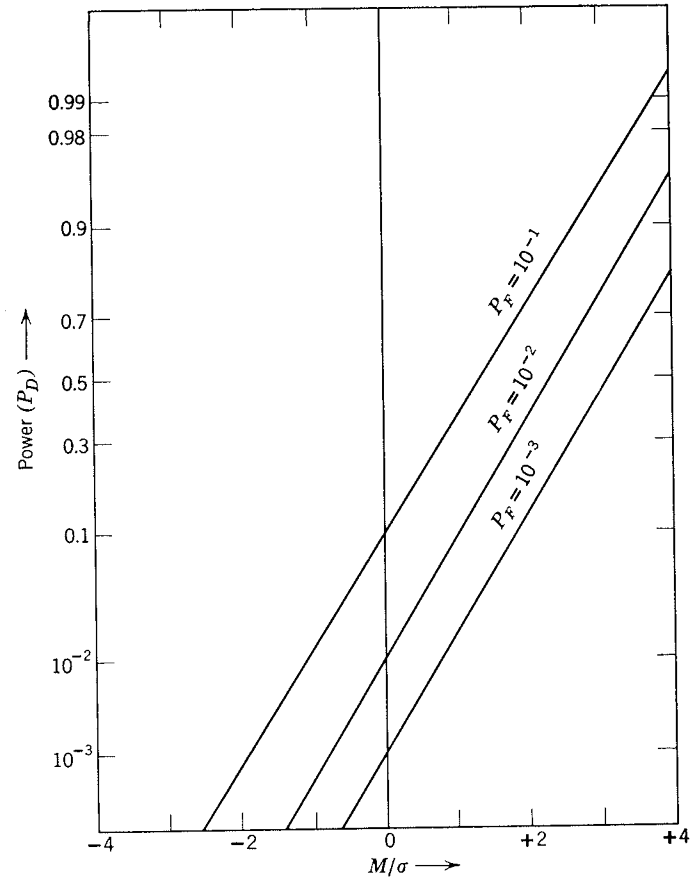


Fig. 2.29 Performance of LRT assuming positive M.

Example 2. The basic probabilities are the same as in Example 1. Once again, $\boldsymbol{\theta} = M$. Instead of one, we have N independent observations, which we denote by the vector $\mathbf{R}$. The probability densities are,

$$p_{\mathbf{r}|m,H_1}(\mathbf{R}|M,H_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(R_i - M)^2}{2\sigma^2}\right),$$
$$p_{\mathbf{r}|m,H_0}(R|M,H_0) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R_i^2}{2\sigma^2}\right). \tag{306}$$

In this example H_1 is a composite hypothesis and H_0, a simple hypothesis. From (198)

$$\hat{M}_1 = \frac{1}{N}\sum_{i=1}^{N} R_i. \tag{307}$$

Then

$$\Lambda_g(\mathbf{R}) = \frac{\displaystyle\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{[R_i - (1/N)\sum_{j=1}^{N} R_j]^2}{2\sigma^2}\right\}}{\displaystyle\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-R_i^2/2\sigma^2\right)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{308}$$

Canceling common terms and taking the logarithm, we have

$$\ln \Lambda_g(\mathbf{R}) = \frac{1}{2\sigma^2 N}\left(\sum_{i=1}^{N} R_i\right)^2 \underset{H_0}{\overset{H_1}{\gtrless}} \ln \gamma. \tag{309}$$

The left side of (309) is always greater than or equal to zero. Thus, γ can always be chosen greater than or equal to one. Therefore, an equivalent test is

$$\left(\frac{1}{N^{1/2}}\sum_{i=1}^{N} R_i\right)^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_1^2 \tag{310}$$

where $\gamma_1 \geq 0$. Equivalently,

$$|z| \triangleq \left|\frac{1}{N^{1/2}}\sum_{i=1}^{N} R_i\right| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_1. \tag{311}$$

The power function of this test follows easily. The variable z has a variance equal

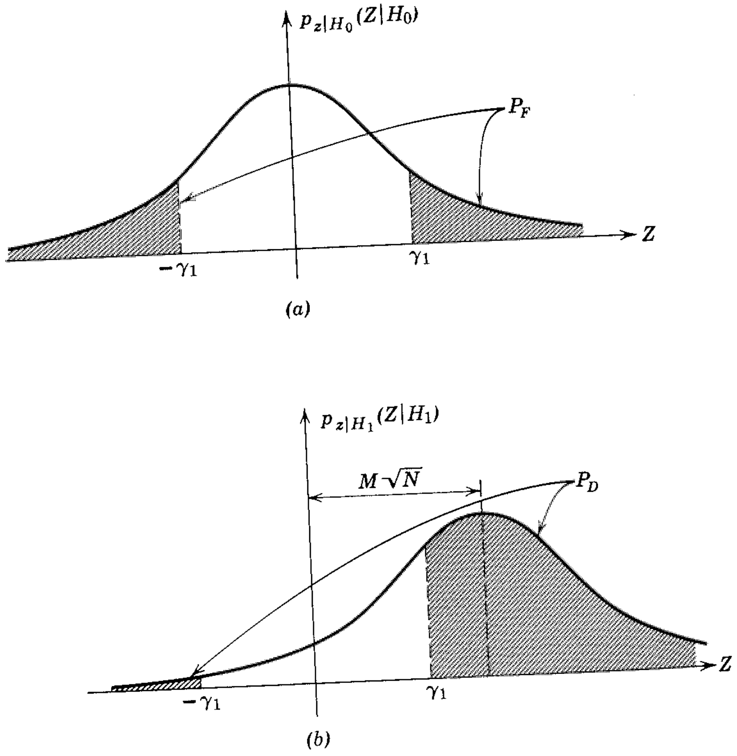


Fig. 2.30 Errors in generalized likelihood ratio test: [a] P_F calculation; [b] P_D calculation.

to σ^2. On H_0 its mean is zero and on H_1 its mean is $M\sqrt{N}$. The densities are sketched in Fig. 2.30.

$$\begin{aligned} P_F &= \int_{-\infty}^{-\gamma_1} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{Z^2}{2\sigma^2}\right) dZ + \int_{\gamma_1}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{Z^2}{2\sigma^2}\right) dZ \\ &= 2\,\text{erfc}_*\left(\frac{\gamma_1}{\sigma}\right) \end{aligned} \tag{312}$$

and

$$\begin{aligned} P_D(M) &= \int_{-\infty}^{-\gamma_1} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(Z - M\sqrt{N})^2}{2\sigma^2}\right] dZ \\ &\quad + \int_{\gamma_1}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(Z - M\sqrt{N})^2}{2\sigma^2}\right] dZ \\ &= \text{erfc}_*\left[\frac{\gamma_1 + M\sqrt{N}}{\sigma}\right] + \text{erfc}_*\left[\frac{\gamma_1 - M\sqrt{N}}{\sigma}\right]. \end{aligned} \tag{313}$$

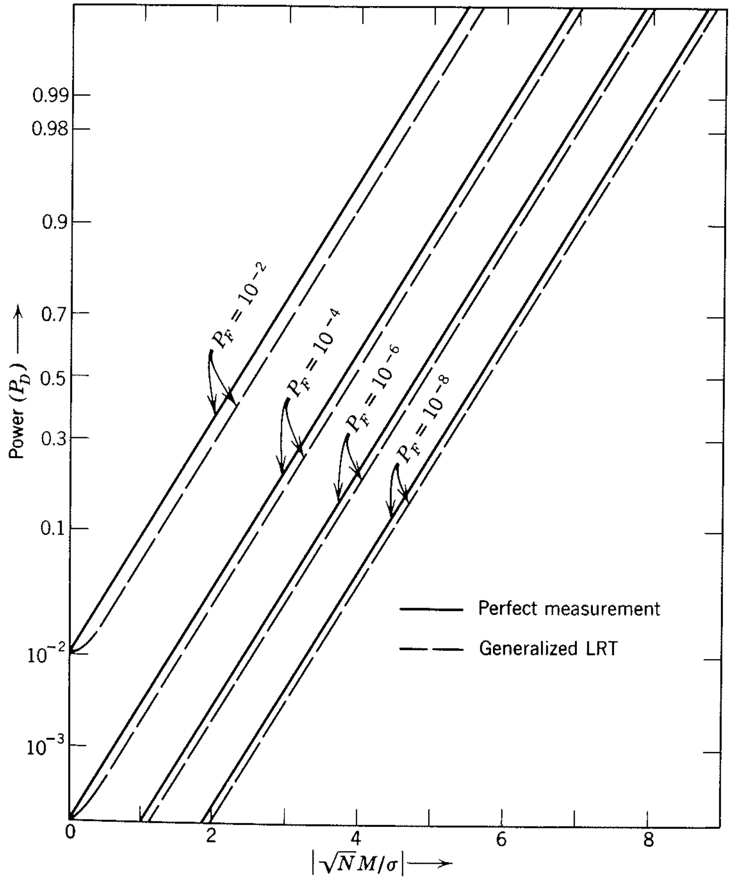


Fig. 2.31 Power function: generalized likelihood ratio tests.

The resulting power function is plotted in Fig. 2.31. The perfect measurement bound is shown for comparison purposes. As we would expect from our discussion of ML estimates, the difference approaches zero as $\sqrt{N}\, M/\sigma \to \infty$.

Just as there are cases in which the ML estimates give poor results, there are others in which the generalized likelihood ratio test may give bad results. In these cases we must look for other test procedures. Fortunately, in most of the physical problems of interest to us either a UMP test will exist or a generalized likelihood ratio test will give satisfactory results.

2.6 THE GENERAL GAUSSIAN PROBLEM

All of our discussion up to this point has dealt with arbitrary probability densities. In the binary detection case $p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)$ and $p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)$ were not constrained to have any particular form. Similarly, in the estimation problem $p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})$ was not constrained. In the classical case, constraints are not particularly necessary. When we begin our discussion of the waveform problem, we shall find that most of our discussions concentrate on problems in which the conditional density of $\mathbf{r}$ is Gaussian. We discuss this class of problem in detail in this section. The material in this section and the problems associated with it lay the groundwork for many of the results in the sequel. We begin by defining a Gaussian random vector and the general Gaussian problem.

Definition. A set of random variables $r_1, r_2, \ldots, r_N$ are defined as jointly Gaussian if all their linear combinations are Gaussian random variables.

Definition. A vector $\mathbf{r}$ is a Gaussian random vector when its components $r_1, r_2, \ldots, r_N$ are jointly Gaussian random variables.

In other words, if

$$z = \sum_{i=1}^{N} g_i r_i \triangleq \mathbf{G}^T\mathbf{r} \tag{314}$$

is a Gaussian random variable for all finite $\mathbf{G}^T$, then $\mathbf{r}$ is a Gaussian vector.

If we define

$$E(\mathbf{r}) = \mathbf{m} \tag{315}$$

and

$$\operatorname{Cov}(\mathbf{r}) = E[(\mathbf{r} - \mathbf{m})(\mathbf{r}^T - \mathbf{m}^T)] \triangleq \mathbf{\Lambda}, \tag{316}$$

then (314) implies that the characteristic function of $\mathbf{r}$ is

$$M_{\mathbf{r}}(j\mathbf{v}) \triangleq E[e^{j\mathbf{v}^T\mathbf{r}}] = \exp(+j\mathbf{v}^T\mathbf{m} - \tfrac{1}{2}\mathbf{v}^T\mathbf{\Lambda}\mathbf{v}) \tag{317}$$

and assuming $\mathbf{\Lambda}$ is nonsingular the probability density of $\mathbf{r}$ is

$$p_{\mathbf{r}}(\mathbf{R}) = [(2\pi)^{N/2}|\mathbf{\Lambda}|^{1/2}]^{-1} \exp[-\tfrac{1}{2}(\mathbf{R}^T - \mathbf{m}^T)\mathbf{\Lambda}^{-1}(\mathbf{R} - \mathbf{m})]. \tag{318}$$

The proof is straightforward (e.g., Problem 2.6.20).

Definition. A hypothesis testing problem is called a general Gaussian problem if $p_{\mathbf{r}|H_i}(\mathbf{R}|H_i)$ is a Gaussian density on all hypotheses. An estimation problem is called a general Gaussian problem if $p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})$ has a Gaussian density for all $\mathbf{A}$.

We discuss the binary hypothesis testing version of the general Gaussian problem in detail in the text. The M-hypothesis and the estimation problems are developed in the problems. The basic model for the binary detection problem is straightforward. We assume that the observation space is N-dimensional. Points in the space are denoted by the N-dimensional vector (or column matrix) $\mathbf{r}$:

$$\mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}. \tag{319}$$

Under the first hypothesis H_1 we assume that $\mathbf{r}$ is a Gaussian random vector, which is completely specified by its mean vector and covariance matrix. We denote these quantities as

$$E[\mathbf{r}|H_1] = \begin{bmatrix} E(r_1|H_1) \\ E(r_2|H_1) \\ \vdots \\ E(r_N|H_1) \end{bmatrix} \triangleq \begin{bmatrix} m_{11} \\ m_{12} \\ \vdots \\ m_{1N} \end{bmatrix} \triangleq \mathbf{m}_1. \tag{320}$$

The covariance matrix is

$$\mathbf{K}_1 \triangleq E[(\mathbf{r} - \mathbf{m}_1)(\mathbf{r}^T - \mathbf{m}_1{}^T)|H_1]$$

$$= \begin{bmatrix} {}_1K_{11} & {}_1K_{12} & {}_1K_{13} & \cdots & {}_1K_{1N} \\ {}_1K_{21} & {}_1K_{22} & \ddots & & \\ \vdots & \vdots & & & \vdots \\ {}_1K_{N1} & & & & {}_1K_{NN} \end{bmatrix}. \tag{321}$$

We define the inverse of $\mathbf{K}_1$ as $\mathbf{Q}_1$

$$\mathbf{Q}_1 \triangleq \mathbf{K}_1^{-1} \tag{322}$$

$$\mathbf{Q}_1\mathbf{K}_1 = \mathbf{K}_1\mathbf{Q}_1 = \mathbf{I}, \tag{323}$$

where **I** is the identity matrix (ones on the diagonal and zeroes elsewhere). Using (320), (321), (322), and (318), we may write the probability density of **r** on H_1,

$$p_{\mathbf{r}|H_1}(\mathbf{R}|H_1) = [(2\pi)^{N/2}|\mathbf{K}_1|^{1/2}]^{-1}\exp\left[-\tfrac{1}{2}(\mathbf{R}^T - \mathbf{m}_1^T)\mathbf{Q}_1(\mathbf{R} - \mathbf{m}_1)\right]. \tag{324}$$

Going through a similar set of definitions for H_0, we obtain the probability density

$$p_{\mathbf{r}|H_0}(\mathbf{R}|H_0) = [(2\pi)^{N/2}|\mathbf{K}_0|^{1/2}]^{-1}\exp\left[-\tfrac{1}{2}(\mathbf{R}^T - \mathbf{m}_0^T)\mathbf{Q}_0(\mathbf{R} - \mathbf{m}_0)\right]. \tag{325}$$

Using the definition in (13), the likelihood ratio test follows easily:

$$\Lambda(\mathbf{R}) \triangleq \frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)} = \frac{|\mathbf{K}_0|^{1/2}\exp\left[-\frac{1}{2}(\mathbf{R}^T - \mathbf{m}_1^T)\mathbf{Q}_1(\mathbf{R} - \mathbf{m}_1)\right]}{|\mathbf{K}_1|^{1/2}\exp\left[-\frac{1}{2}(\mathbf{R}^T - \mathbf{m}_0^T)\mathbf{Q}_0(\mathbf{R} - \mathbf{m}_0)\right]} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{326}$$

Taking logarithms, we obtain

$$\boxed{\begin{aligned}\tfrac{1}{2}(\mathbf{R}^T - \mathbf{m}_0^T)\,\mathbf{Q}_0(\mathbf{R} - \mathbf{m}_0) - \tfrac{1}{2}(\mathbf{R}^T - \mathbf{m}_1^T)\,\mathbf{Q}_1(\mathbf{R} - \mathbf{m}_1)&\\ \underset{H_0}{\overset{H_1}{\gtrless}} \ln\eta + \tfrac{1}{2}\ln|\mathbf{K}_1| - \tfrac{1}{2}\ln|\mathbf{K}_0| \triangleq \gamma^*.&\end{aligned}} \tag{327}$$

We see that the test consists of finding the difference between two *quadratic forms.* The result in (327) is basic to many of our later discussions. For this reason we treat various cases of the general Gaussian problem in some detail. We begin with the simplest.

2.6.1 Equal Covariance Matrices

The first special case of interest is the one in which the covariance matrices on the two hypotheses are equal,

$$\mathbf{K}_1 = \mathbf{K}_0 \triangleq \mathbf{K}, \tag{328}$$

but the means are different.

Denote the inverse as **Q**:

$$\mathbf{Q} = \mathbf{K}^{-1}. \tag{329}$$

Substituting into (327), multiplying the matrices, canceling common terms, and using the symmetry of **Q**, we have

$$(\mathbf{m}_1^T - \mathbf{m}_0^T)\mathbf{Q}\mathbf{R} \underset{H_0}{\overset{H_1}{\gtrless}} \ln\eta + \tfrac{1}{2}(\mathbf{m}_1^T\mathbf{Q}\mathbf{m}_1 - \mathbf{m}_0^T\mathbf{Q}\mathbf{m}_0) \triangleq \gamma'_*. \tag{330}$$

We can simplify this expression by defining a vector corresponding to the difference in the mean value vectors on the two hypotheses:

$$\Delta\mathbf{m} \triangleq \mathbf{m}_1 - \mathbf{m}_0. \tag{331}$$

Then (327) becomes

$$\boxed{l(\mathbf{R}) \triangleq \Delta\mathbf{m}^T\mathbf{Q}\mathbf{R} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'_*} \tag{332}$$

or, equivalently,

$$l(\mathbf{R}) \triangleq \mathbf{R}^T\mathbf{Q}\,\Delta\mathbf{m} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'_*. \tag{333}$$

The quantity on the left is a *scalar* Gaussian random variable, for it was obtained by a linear transformation of jointly Gaussian random variables. Therefore, as we discussed in Example 1 on pp. 36–38, we can completely characterize the performance of the test by the quantity d^2. In that example, we defined d as the distance between the means on the two hypothesis when the variance was normalized to equal one. An identical definition is,

$$d^2 \triangleq \frac{[E(l|H_1) - E(l|H_0)]^2}{\text{Var}\,(l|H_0)}. \tag{334}$$

Substituting (320) into the definition of l, we have

$$E(l|H_1) = \Delta\mathbf{m}^T\mathbf{Q}\mathbf{m}_1 \tag{335}$$

and

$$E(l|H_0) = \Delta\mathbf{m}^T\mathbf{Q}\mathbf{m}_0. \tag{336}$$

Using (332), (333), and (336) we have

$$\text{Var}\,[l|H_0] = E\{[\Delta\mathbf{m}^T\mathbf{Q}(\mathbf{R} - \mathbf{m}_0)][(\mathbf{R}^T - \mathbf{m}_0^T)\mathbf{Q}\,\Delta\mathbf{m}]\}. \tag{337}$$

Using (321) to evaluate the expectation and then (323), we have

$$\text{Var}\,[l|H_0] = \Delta\mathbf{m}^T\mathbf{Q}\,\Delta\mathbf{m}. \tag{338}$$

Substituting (335), (336), and (338) into (334), we obtain

$$\boxed{d^2 = \Delta\mathbf{m}^T\mathbf{Q}\,\Delta\mathbf{m}}. \tag{339}$$

Thus the performance for the equal covariance Gaussian case is completely determined by the quadratic form in (339). We now interpret it for some cases of interest.

Case 1. Independent Components with Equal Variance. Each r_i has the same variance σ^2 and is statistically independent. Thus

$$\mathbf{K} = \sigma^2\mathbf{I} \tag{340}$$

and

$$\mathbf{Q} = \frac{1}{\sigma^2}\mathbf{I}. \tag{341}$$

Substituting (341) into (339), we obtain

$$d^2 = \mathbf{\Delta m}^T \frac{1}{\sigma^2} \mathbf{I}\, \mathbf{\Delta m} = \frac{1}{\sigma^2} \mathbf{\Delta m}^T\, \mathbf{\Delta m} = \frac{1}{\sigma^2} |\mathbf{\Delta m}|^2 \tag{342}$$

or

$$\boxed{d = \frac{|\mathbf{\Delta m}|}{\sigma}}. \tag{343}$$

We see that d corresponds to the *distance* between the two mean-value vectors divided by the standard deviation of R_i. This is shown in Fig. 2.32. In (332) we see that

$$l = \frac{1}{\sigma^2} \mathbf{\Delta m}^T \mathbf{R}. \tag{344}$$

Thus the sufficient statistic is just the dot (or scalar) product of the observed vector $\mathbf{R}$ and the mean difference vector $\mathbf{\Delta m}$.

Case 2. Independent Components with Unequal Variances. Here the r_i are statistically independent but have unequal variances. Thus

$$\mathbf{K} = \begin{bmatrix} {\sigma_1}^2 & & & 0 \\ & {\sigma_2}^2 & & \\ & & \ddots & \\ 0 & & & {\sigma_N}^2 \end{bmatrix} \tag{345}$$

and

$$\mathbf{Q} = \begin{bmatrix} \frac{1}{{\sigma_1}^2} & & & 0 \\ & \frac{1}{{\sigma_2}^2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{{\sigma_N}^2} \end{bmatrix}. \tag{346}$$

Substituting into (339) and performing the multiplication, we have

$$d^2 = \sum_{i=1}^{N} \frac{(\Delta m_i)^2}{{\sigma_i}^2}. \tag{347}$$

Now the various difference components contribute to d^2 with weighting that is inversely proportional to the variance along that coordinate. We can also interpret the result as distance in a new coordinate system.

Let

$$\mathbf{\Delta m}' = \begin{bmatrix} \frac{1}{\sigma_1} m_1 \\ \frac{1}{\sigma_2} m_2 \\ \vdots \\ \frac{1}{\sigma_N} m_N \end{bmatrix} \tag{348}$$

and

$$R_i' = \frac{1}{\sigma_i} R_i. \tag{349}$$

This transformation changes the scale on each axis so that the variances are all equal to one. We see that d corresponds exactly to the difference vector in this "scaled" coordinate system.

The sufficient statistic is

$$l(\mathbf{R}) = \sum_{i=1}^{N} \frac{\Delta m_i \cdot R_i}{{\sigma_i}^2}. \tag{350}$$

In the scaled coordinate system it is the dot product of the two vectors

$$l(\mathbf{R}') = \mathbf{\Delta m}'^T \mathbf{R}'. \tag{351}$$

Case 3. This is the general case. A satisfactory answer for l and d is already available in (332) and (339):

$$l(\mathbf{R}) = \mathbf{\Delta m}^T \mathbf{Q} \mathbf{R} \tag{352}$$

and

$$d^2 = \mathbf{\Delta m}^T \mathbf{Q}\, \mathbf{\Delta m}. \tag{353}$$

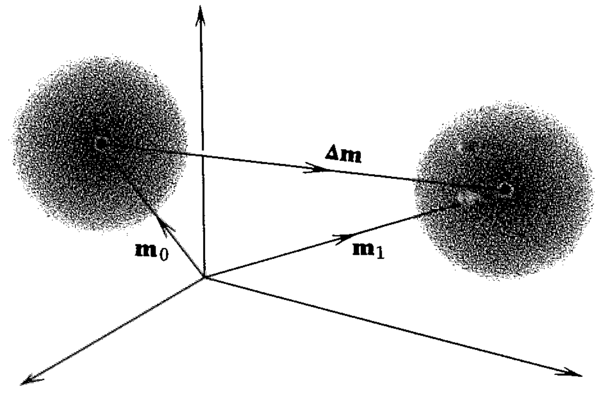


Fig. 2.32 Mean-value vectors.

Valuable insight into the important features of the problem can be gained by looking at it in a different manner.

The key to the simplicity in Cases 1 and 2 is the diagonal covariance matrix. This suggests that we try to represent $\mathbf{R}$ in a new coordinate system in which the components are statistically independent random variables. In Fig. 2.33*a* we show the observation in the original coordinate system. In Fig. 2.33*b* we show a new set of coordinate axes, which we denote by the orthogonal unit vectors $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \dots, \boldsymbol{\phi}_N$:

$$\boldsymbol{\phi}_i^T \boldsymbol{\phi}_j = \delta_{ij}. \tag{354}$$

We denote the observation in the new coordinate system by $\mathbf{r}'$. We want to choose the orientation of the new system so that the components r_i' and r_j' are uncorrelated (and therefore statistically independent, for they are Gaussian) for all $i \neq j$. In other words,

$$E[(r_i' - m_i')(r_j' - m_j')] = \lambda_i \delta_{ij}, \tag{355}$$

where

$$m_i' \triangleq E(r_i') \tag{356}$$

and

$$\text{Var}\,[r_i'] \triangleq \lambda_i. \tag{357}$$

Now the components of $\mathbf{r}'$ can be expressed simply in terms of the dot product of the original vector $\mathbf{r}$ and the various unit vectors

$$r_i' = \mathbf{r}^T \boldsymbol{\phi}_i = \boldsymbol{\phi}_i^T \mathbf{r}. \tag{358}$$

Using (358) in (355), we obtain

$$E[\boldsymbol{\phi}_i^T(\mathbf{r} - \mathbf{m})(\mathbf{r}^T - \mathbf{m}^T)\boldsymbol{\phi}_j] = \lambda_i \delta_{ij}. \tag{359}$$

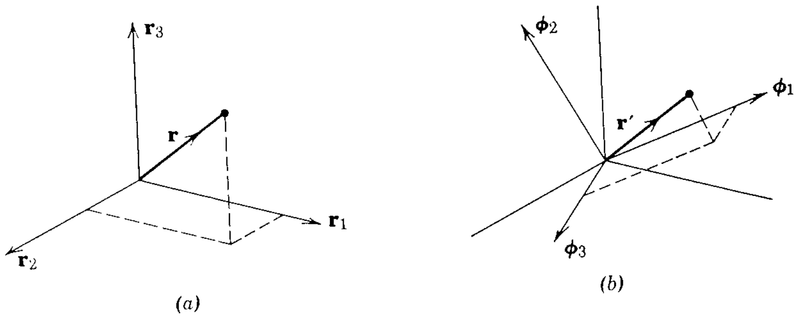


Fig. 2.33 Coordinate systems: [*a*] original coordinate system; [*b*] new coordinate system.

The expectation of the random part is just $\mathbf{K}$ [see (321)]. Therefore (359) becomes

$$\lambda_i \delta_{ij} = \boldsymbol{\phi}_i^T \mathbf{K} \boldsymbol{\phi}_j. \tag{360}$$

This will be satisfied if and only if

$$\lambda_j \boldsymbol{\phi}_j = \mathbf{K} \boldsymbol{\phi}_j \qquad \text{for } j = 1, 2, \dots, N. \tag{361}$$

To check the "if" part of this result, substitute (361) into (360):

$$\lambda_i \delta_{ij} = \boldsymbol{\phi}_i^T \lambda_j \boldsymbol{\phi}_j = \lambda_j \delta_{ij}, \tag{362}$$

where the right equality follows from (354). The "only if" part follows using a simple proof by contradiction. Now (361) can be written with the j subscript suppressed:

$$\boxed{\lambda \boldsymbol{\phi} = \mathbf{K} \boldsymbol{\phi}.} \tag{363}$$

We see that the question of finding the proper coordinate system reduces to the question of whether we can find N solutions to (363) that satisfy (354).

It is instructive to write (363) out in detail. Each $\boldsymbol{\phi}$ is a vector with N components:

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \vdots \\ \phi_N \end{bmatrix}. \tag{364}$$

Substituting (364) into (363), we have

$$\begin{aligned} K_{11}\phi_1 + K_{12}\phi_2 + \cdots + K_{1N}\phi_N &= \lambda\phi_1 \\ K_{21}\phi_1 + K_{22}\phi_2 + \cdots + K_{2N}\phi_N &= \lambda\phi_2 \\ &\vdots \\ K_{N1}\phi_1 + K_{N2}\phi_2 + \cdots + K_{NN}\phi_N &= \lambda\phi_N \end{aligned} \tag{365}$$

We see that (365) corresponds to a set of N homogeneous simultaneous equations. A nontrivial solution will exist if and only if the determinant of the coefficient matrix is zero. In other words, if and only if

$$|\mathbf{K} - \lambda \mathbf{I}| = \begin{vmatrix} K_{11} - \lambda & K_{12} & K_{13} & \cdots \\ K_{21} & K_{22} - \lambda & K_{23} & \\ K_{31} & K_{32} & \ddots & \\ \vdots & \vdots & & K_{NN} - \lambda \end{vmatrix} = 0. \tag{366}$$

We see that this is an Nth-order polynomial in λ. The N roots, denoted by $\lambda_1, \lambda_2, \ldots, \lambda_N$, are called the *eigenvalues* of the covariance matrix $\mathbf{K}$. It can be shown that the following properties are true (e.g., [16] or [18]):

1. Because $\mathbf{K}$ is symmetric, the eigenvalues are real.
2. Because $\mathbf{K}$ is a covariance matrix, the eigenvalues are nonnegative. (Otherwise we would have random variables with negative variances.)

For each λ_i we can find a solution $\boldsymbol{\phi}_i$ to (363). Because there is an arbitrary constant associated with each solution to (363), we may choose the $\boldsymbol{\phi}_i$ to have unit length

$$\boldsymbol{\phi}_i^T \boldsymbol{\phi}_i = 1. \tag{367}$$

These solutions are called the normalized *eigenvectors* of $\mathbf{K}$. Two other properties may also be shown for symmetric matrices.

3. If the roots λ_i are distinct, the corresponding eigenvectors are orthogonal.
4. If a particular root λ_j is of multiplicity M, the M associated eigenvectors are linearly independent. They can be chosen to be orthonormal.

We have now described a coordinate system in which the observations are statistically independent. The mean difference vector can be expressed as

$$\begin{aligned}
\Delta m_1' &= \boldsymbol{\phi}_1^T\, \Delta\mathbf{m} \\
\Delta m_2' &= \boldsymbol{\phi}_2^T\, \Delta\mathbf{m} \\
&\;\vdots \\
\Delta m_N' &= \boldsymbol{\phi}_N^T\, \Delta\mathbf{m}
\end{aligned} \tag{368}$$

or in vector notation

$$\Delta\mathbf{m}' = \begin{bmatrix} \boldsymbol{\phi}_1^T \\ \hdashline \boldsymbol{\phi}_2^T \\ \hdashline \vdots \\ \hdashline \boldsymbol{\phi}_N^T \end{bmatrix} \Delta\mathbf{m} \triangleq \mathbf{W}\, \Delta\mathbf{m}. \tag{369}$$

The resulting sufficient statistic in the new coordinate system is

$$l(\mathbf{R}') = \sum_{i=1}^{N} \frac{\Delta m_i' \cdot R_i'}{\lambda_i} \tag{370}$$

and d^2 is

$$\boxed{d^2 = \sum_{i=1}^{N} \frac{(\Delta m_i')^2}{\lambda_i}.} \tag{371}$$

The derivation leading to (371) has been somewhat involved, but the result is of fundamental importance, for it demonstrates that there always exists a coordinate system in which the random variables are uncorrelated and that the new system is related to the old system by a linear transformation. To illustrate the technique we consider a simple example.

Example. For simplicity we let $N = 2$ and $\mathbf{m}_0 = \mathbf{0}$. Let

$$\mathbf{K} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{372}$$

and

$$\mathbf{m}_1 = \begin{bmatrix} m_{11} \\ m_{12} \end{bmatrix}. \tag{373}$$

To find the eigenvalues we solve

$$\begin{vmatrix} 1-\lambda & \rho \\ \rho & 1-\lambda \end{vmatrix} = 0 \tag{374}$$

or

$$(1 - \lambda^2) - \rho^2 = 0. \tag{375}$$

Solving,

$$\begin{aligned} \lambda_1 &= 1 + \rho, \\ \lambda_2 &= 1 - \rho. \end{aligned} \tag{376}$$

To find $\boldsymbol{\phi}_1$ we substitute λ_1 into (365),

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \phi_{11} \\ \phi_{12} \end{bmatrix} = \begin{bmatrix} (1+\rho)\phi_{11} \\ (1+\rho)\phi_{12} \end{bmatrix} \tag{377}$$

Solving, we obtain

$$\phi_{11} = \phi_{12}. \tag{378}$$

Normalizing gives

$$\boldsymbol{\phi}_1 = \begin{bmatrix} +\dfrac{1}{\sqrt{2}} \\ +\dfrac{1}{\sqrt{2}} \end{bmatrix}. \tag{379}$$

Similarly,

$$\boldsymbol{\phi}_2 = \begin{bmatrix} +\dfrac{1}{\sqrt{2}} \\ -\dfrac{1}{\sqrt{2}} \end{bmatrix}. \tag{380}$$

The old and new axes are shown in Fig. 2.34. The transformation is

$$\mathbf{W} = \left[\begin{array}{c:c} +\dfrac{1}{\sqrt{2}} & +\dfrac{1}{\sqrt{2}} \\ \hdashline +\dfrac{1}{\sqrt{2}} & -\dfrac{1}{\sqrt{2}} \end{array}\right] \tag{381}$$

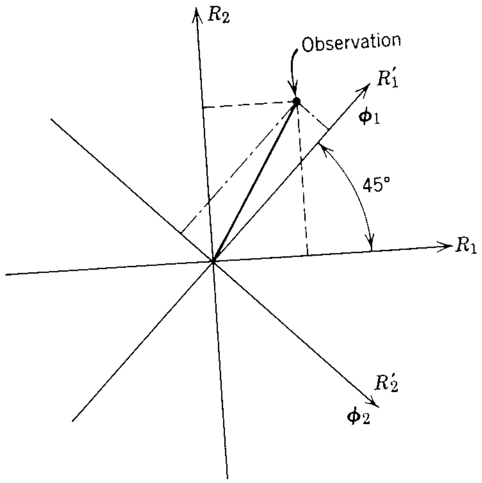


Fig. 2.34 Rotation of axes.

$$R_1' = \frac{R_1 + R_2}{\sqrt{2}},$$
$$R_2' = \frac{R_1 - R_2}{\sqrt{2}},$$
$$m_{11}' = \frac{m_{11} + m_{12}}{\sqrt{2}},$$
$$m_{12}' = \frac{m_{11} - m_{12}}{\sqrt{2}}. \tag{382}$$

The sufficient statistic is obtained by using (382) in (370),

$$l(\mathbf{R}') = \frac{1}{1+\rho}\frac{(R_1 + R_2)(m_{11} + m_{12})}{2} + \frac{1}{1-\rho}\frac{(R_1 - R_2)(m_{11} - m_{12})}{2} \tag{383}$$

and d^2 is

$$d^2 = \frac{(m_{11} + m_{12})^2}{2(1+\rho)} + \frac{(m_{11} - m_{12})^2}{2(1-\rho)} = \frac{(m_{11}')^2}{(1+\rho)} + \frac{(m_{12}')^2}{(1-\rho)}. \tag{384}$$

To illustrate a typical application in which the transformation is important we consider a simple optimization problem. The length of the mean vector is constrained,

$$|\mathbf{m}_1|^2 = 1. \tag{385}$$

We want to choose m_{11} and m_{12} to maximize d^2. Because our transformation is a rotation, it preserves lengths

$$|\mathbf{m}_1'|^2 = 1. \tag{386}$$

Looking at (384), we obtain the solution by inspection:

If $\rho > 0$, choose $m_{11}' = 0$ and $m_{12}' = 1$.

If $\rho < 0$, choose $m_{11}' = 1$ and $m_{12}' = 0$.

If $\rho = 0$, all vectors satisfying (385) give the same d^2.

We see that this corresponds to choosing the mean-value vector to be equal to the eigenvector with the smallest eigenvalue. This result can be easily extended to N dimensions.

The result in this example is characteristic of a wide class of optimization problems in which the solution corresponds to an eigenvector (or the waveform analog to it).

In this section, we have demonstrated that when the covariance matrices on the two hypotheses are equal the sufficient statistic $l(\mathbf{R})$ is a Gaussian random variable obtained by a linear transformation of $\mathbf{R}$. The performance for any threshold setting is determined by using the value of d^2 given by (339) on the ROC in Fig. 2.9. Because the performance improves monotonically with increasing d^2, we can use any freedom in the parameters to maximize d^2 without considering the ROC explicitly.

2.6.2 Equal Mean Vectors

In the second special case of interest the mean-value vectors on the two hypotheses are equal. In other words,

$$\mathbf{m}_1 = \mathbf{m}_0 \triangleq \mathbf{m}. \tag{387}$$

Substituting (387) into (327), we have

$$\tfrac{1}{2}(\mathbf{R}^T - \mathbf{m}^T)(\mathbf{Q}_0 - \mathbf{Q}_1)(\mathbf{R} - \mathbf{m}) \underset{H_0}{\overset{H_1}{\gtrless}} \ln \eta + \tfrac{1}{2}\ln\frac{|\mathbf{K}_1|}{|\mathbf{K}_0|} = \gamma^*. \tag{388}$$

Because the mean-value vectors contain no information that will tell us which hypothesis is true, the likelihood test subtracts them from the received vector. Therefore, without loss of generality, we may assume that $\mathbf{m} = \mathbf{0}$.

We denote the difference of the inverse matrices as $\Delta\mathbf{Q}$:

$$\Delta\mathbf{Q} \triangleq \mathbf{Q}_0 - \mathbf{Q}_1. \tag{389}$$

The likelihood ratio test may be written as

$$\boxed{l(\mathbf{R}) \triangleq \mathbf{R}^T\,\Delta\mathbf{Q}\mathbf{R} \underset{H_0}{\overset{H_1}{\gtrless}} 2\gamma^* \triangleq \gamma'.} \tag{390}$$

Note that $l(\mathbf{R})$ is the dot product of two Gaussian vectors, $\mathbf{R}^T$ and $\Delta\mathbf{QR}$. Thus, $l(\mathbf{R})$ is not a Gaussian random variable.

We now consider the behavior of this test for some interesting special cases.

Case 1. Diagonal Covariance Matrix on H_0: Equal Variances. Here the R_i on H_0 are statistically independent variables with equal variances:

$$\mathbf{K}_0 = \sigma_n{}^2\mathbf{I}. \tag{391}$$

We shall see later that (391) corresponds to the physical situation in which there is "noise" only on H_0. The following notation is convenient:

$$r_i = n_i, \qquad H_0. \tag{392}$$

On H_1 the r_i contain the same variable as on H_0, plus additional signal components that may be correlated:

$$r_i = s_i + n_i, \qquad H_1, \tag{393}$$

$$\mathbf{K}_1 = \mathbf{K}_s + \sigma_n^2\mathbf{I},$$

where the matrix $\mathbf{K}_s$ represents the covariance matrix of the signal components. Then

$$\mathbf{Q}_0 = \frac{1}{\sigma_n^2}\mathbf{I} \tag{394}$$

and

$$\mathbf{Q}_1 = \frac{1}{\sigma_n^2}\left(\mathbf{I} + \frac{1}{\sigma_n^2}\mathbf{K}_s\right)^{-1}. \tag{395}$$

It is convenient to write (395) as

$$\mathbf{Q}_1 = \frac{1}{\sigma_n^2}[\mathbf{I} - \mathbf{H}], \tag{396}$$

which implies

$$\mathbf{H} = (\sigma_n^2\mathbf{I} + \mathbf{K}_s)^{-1}\mathbf{K}_s = \mathbf{K}_s(\sigma_n^2\mathbf{I} + \mathbf{K}_s)^{-1} = \mathbf{Q}_0 - \mathbf{Q}_1 = \Delta\mathbf{Q}. \tag{397}$$

The $\mathbf{H}$ matrix has an important interpretation which we shall develop later. We take the first expression in (397) as its definition. Substituting (397) into (389) and the result into (390), we have

$$\boxed{l(\mathbf{R}) = \frac{1}{\sigma_n^2}\mathbf{R}^T\mathbf{H}\mathbf{R} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'.} \tag{398}$$

Several subcases are important.

Case 1A. Uncorrelated, Identically Distributed Signal Components. In this case the signal components s_i are independent variables with identical variances:

$$\mathbf{K}_1 = \sigma_s^2\mathbf{I}. \tag{399}$$

Then

$$\mathbf{H} = (\sigma_n^2\mathbf{I} + \sigma_s^2\mathbf{I})^{-1}\sigma_s^2\mathbf{I}, \tag{400}$$

or

$$\mathbf{H} = \frac{\sigma_s^2}{\sigma_n^2 + \sigma_s^2}\mathbf{I} \tag{401}$$

and

$$l(\mathbf{R}) = \frac{1}{\sigma_n^2}\frac{\sigma_s^2}{\sigma_n^2 + \sigma_s^2}\mathbf{R}^T\mathbf{R} = \frac{1}{\sigma_n^2}\frac{\sigma_s^2}{\sigma_n^2 + \sigma_s^2}\sum_{i=1}^{N} R_i^2. \tag{402}$$

The constant can be incorporated in the threshold to give

$$l(\mathbf{R}) \triangleq \sum_{i=1}^{N} R_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma''. \tag{403}$$

We now calculate the performance of the test. On both hypotheses $l(\mathbf{R})$ is the sum of the squares of N Gaussian variables. The difference in the hypotheses is in the variance of the Gaussian variables. For simplicity, we shall assume that N is an even integer.

To find $p_{l|H_0}(L|H_0)$ we observe that the characteristic function of each R_i^2 is

$$\begin{aligned} M_{R_i^2|H_0}(jv) &\triangleq \int_{-\infty}^{\infty} e^{jvR_i^2}\frac{1}{\sqrt{2\pi}\,\sigma_n}e^{-R_i^2/2\sigma_n^2}\,dR_i \\ &= (1 - 2jv\sigma_n^2)^{-\frac{1}{2}}. \end{aligned} \tag{404}$$

Because of the independence of the variables, $M_{l|H_0}(jv)$ can be written as a product. Therefore

$$M_{l|H_0}(jv) = (1 - 2jv\sigma_n^2)^{-N/2}. \tag{405}$$

Taking the inverse transform, we obtain $p_{l|H_0}(L|H_0)$:

$$\begin{aligned} p_{l|H_0}(L|H_0) &= \frac{L^{N/2-1}e^{-L/2\sigma_n^2}}{2^{N/2}\sigma_n^N\Gamma\left(\frac{N}{2}\right)}, \qquad L \geq 0, \\ &= 0, \qquad L < 0, \end{aligned} \tag{406}$$

which is familiar as the χ^2 (chi-square) density function with N degrees of freedom. It is tabulated in several references (e.g., [19] or [3]). For $N = 2$ it is easy to check that it is the simple exponential on p. 41. Similarly,

$$\begin{aligned} p_{l|H_1}(L|H_1) &= \frac{L^{N/2-1}e^{-L/2\sigma_1^2}}{2^{N/2}\sigma_1^N\Gamma\left(\frac{N}{2}\right)}, \qquad L \geq 0, \\ &= 0, \qquad L < 0, \end{aligned} \tag{407}$$

where $\sigma_1^2 \triangleq \sigma_s^2 + \sigma_n^2$.

The expressions for P_D and P_F are,

$$P_D = \int_{\gamma''}^{\infty} [2^{N/2}\sigma_1^N\Gamma(N/2)]^{-1}L^{N/2-1}e^{-L/2\sigma_1^2}\,dL \tag{408}$$

and

$$P_F = \int_{\gamma''}^{\infty} [2^{N/2}\sigma_n^N\Gamma(N/2)]^{-1}L^{N/2-1}e^{-L/2\sigma_n^2}\,dL. \tag{409}$$

Construction of the ROC requires an evaluation of the two integrals. We see that for $N = 2$ we have the same problem as Example 2 on p. 41 and (408) and (409) reduce to

$$P_D = \exp\left(-\frac{\gamma''}{2\sigma_1^2}\right), \tag{410}$$

$$P_F = \exp\left(-\frac{\gamma''}{2\sigma_n^2}\right),$$

and

$$P_F = (P_D)^{(1+\sigma_s^2/\sigma_n^2)}. \tag{411}$$

For the general case there are several methods of proceeding. First, let $M = N/2 - 1$ and $\gamma''' = \gamma''/2\sigma_n^2$. Then write

$$P_F = 1 - \int_0^{\gamma'''} \frac{x^M}{M!} e^{-x}\, dx. \tag{412}$$

The integral, called the incomplete Gamma function, has been tabulated by Pearson [21]:

$$I_\Gamma(u, M) \triangleq \int_0^{u\sqrt{M+1}} \frac{x^M}{M!} e^{-x}\, dx, \tag{413}$$

and

$$P_F = 1 - I_\Gamma\left(\frac{\gamma'''}{\sqrt{M+1}}, M\right). \tag{414}$$

These tables are most useful for $P_F \geq 10^{-6}$ and $M \leq 50$. In a second approach we integrate by parts M times. The result is

$$P_F = \exp(-\gamma''') \sum_{k=0}^{M} \frac{(\gamma''')^k}{k!}. \tag{415}$$

For *small* P_F, γ''' is large and we can approximate the series by the last few terms,

$$P_F = \frac{(\gamma''')^M e^{-\gamma'''}}{M!}\left[1 + \frac{M}{\gamma'''} + \frac{M(M-1)}{(\gamma''')^2} + \cdots\right]. \tag{416}$$

Furthermore, we can approximate the bracket as $(1 - M/\gamma''')^{-1}$. This gives

$$P_F \cong \frac{(\gamma''')^M e^{-\gamma'''}}{M!(1 - M/\gamma''')}. \tag{417}$$

A similar expression for P_D follows in which γ''' is replaced by $\gamma^{iv} \triangleq$ $\gamma''/2\sigma_1^2$. The approximate expression in (417) is useful for manual calcula- tion. In actual practice, we use (415) and calculate the ROC numerically. In Fig. 2.35a we have plotted the receiver operating characteristic for some representative values of N and σ_s^2/σ_n^2.

Two particularly interesting curves are those for $N = 8$, $\sigma_s^2/\sigma_n^2 = 1$ and $N = 2$, $\sigma_s^2/\sigma_n^2 = 4$. In both cases the product $N\sigma_s^2/\sigma_n^2 = 8$. We see that when the desired P_F is greater than 0.3, P_D is higher if the available "signal strength" is divided into more components. This suggests that for each P_F

and product $N\sigma_s^2/\sigma_n^2$ there should be an optimum N. In Chapter 4 we shall see that this problem corresponds to optimum diversity in communication systems and the optimum energy per pulse in radar. In Figs. 2.35b and c we have sketched P_M as a function of N for $P_F = 10^{-2}$ and 10^{-4}, respectively, and various $N\sigma_s^2/\sigma_n^2$ products. We discuss the physical implications of these results in Chapter 4.

Case 1B. Independent Signal Components: Unequal Variances. In this case the signal components s_i are independent variables with variances $\sigma_{s_i}^2$:

$$\mathbf{K}_1 = \begin{bmatrix} \sigma_{s_1}^2 & & & 0 \\ & \sigma_{s_2}^2 & & \\ 0 & & \ddots & \\ & & & \sigma_{s_N}^2 \end{bmatrix}. \tag{418}$$

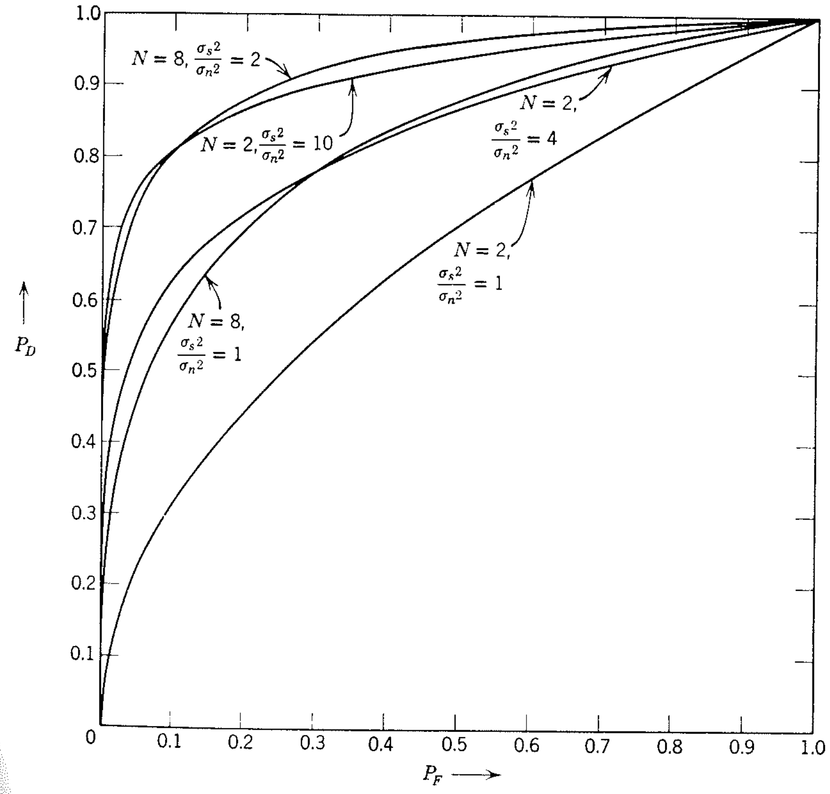


Fig. 2.35 ***a.*** **Receiver operating characteristic: Gaussian variables with identical means and unequal variances on the two hypotheses.**

Then

$$\mathbf{H} = \begin{bmatrix} \dfrac{\sigma_{s_1}{}^2}{\sigma_n{}^2 + \sigma_{s_1}{}^2} & & & 0 \\ & \dfrac{\sigma_{s_2}{}^2}{\sigma_n{}^2 + \sigma_{s_2}{}^2} & & \\ & & \ddots & \\ 0 & & & \dfrac{\sigma_{s_N}{}^2}{\sigma_n{}^2 + \sigma_{s_N}{}^2} \end{bmatrix} \tag{419}$$

and

$$l(\mathbf{R}) = \frac{1}{\sigma_n{}^2} \sum_{i=1}^{N} \frac{\sigma_{s_i}{}^2}{\sigma_n{}^2 + \sigma_{s_i}{}^2} R_i{}^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'. \tag{420}$$

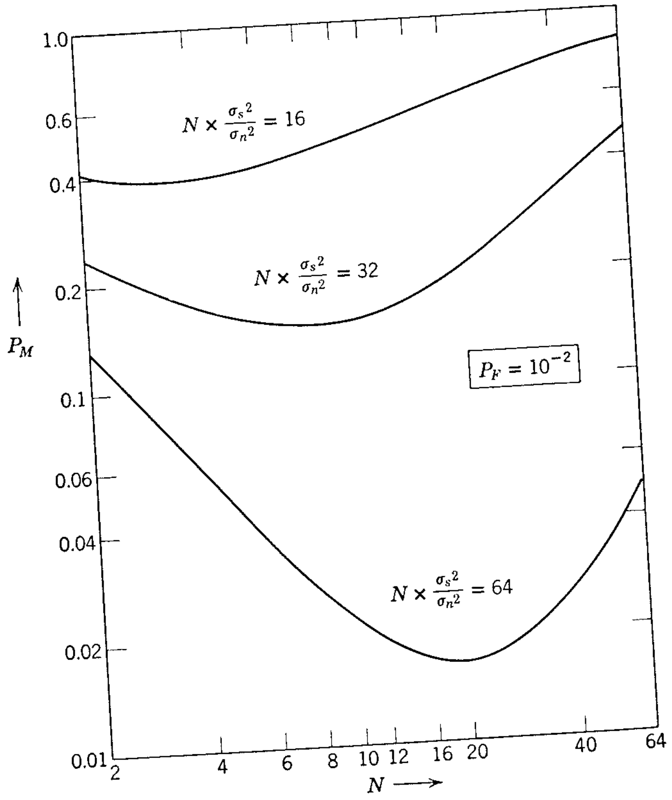


Fig. 2.35 ***b.*** **P_M as a function of N [$P_F = 10^{-2}$].**

The characteristic function of $l(\mathbf{R})$ follows easily, but the calculation of P_F and P_D is difficult. In Section 2.7 we derive approximations to the performance that lead to simpler expressions.

Case 1C. Arbitrary Signal Components. This is, of course, the general case. We revisit it merely to point out that it can always be reduced to Case 1B by an orthogonal transformation (see discussion on pp. 102–106).

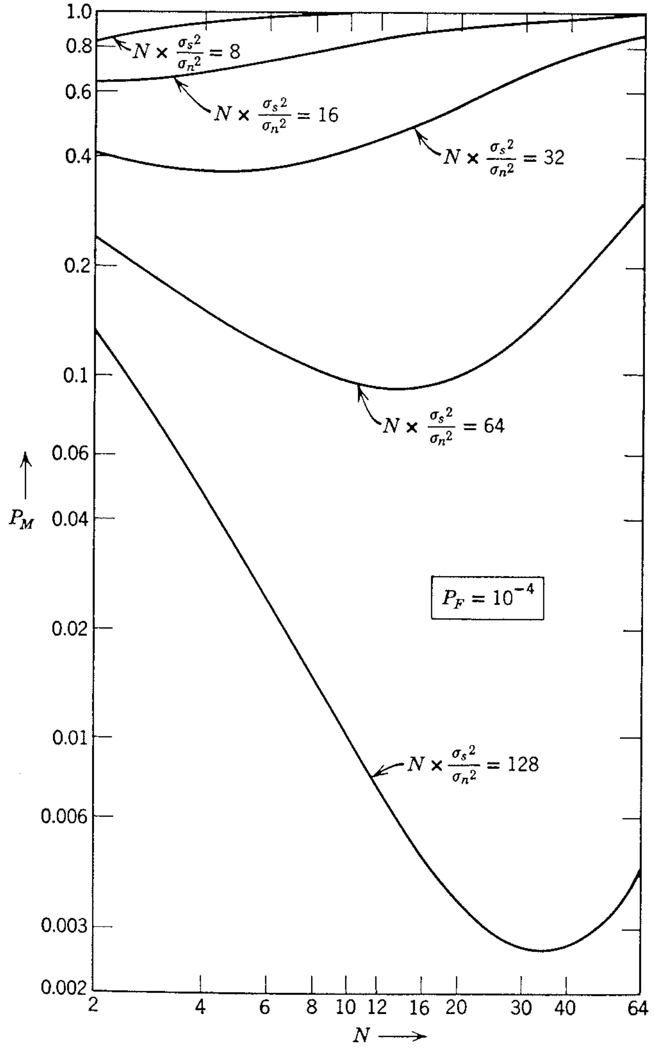


Fig. 2.35 ***c.*** **P_M as a function of N [$P_F = 10^{-4}$].**

Case 2. Symmetric Hypotheses, Uncorrelated Noise. Case 1 was unsymmetric because of the noise-only hypothesis. Here we have the following hypotheses:

$$\begin{aligned} H_1: r_i &= s_i + n_i & i &= 1, \dots, N \\ &\ \ \ \ \ \ n_i & i &= N+1, \dots, 2N, \\ H_0: r_i &= \ \ \ \ \ \ n_i & i &= 1, \dots, N \\ &\ s_i + n_i & i &= N+1, \dots, 2N, \end{aligned} \tag{421}$$

where the n_i are independent variables with variance σ_n^2 and the s_i have a covariance matrix $\mathbf{K}_s$. Then

$$\mathbf{K}_1 = \left[\begin{array}{c|c} \sigma_n^2\mathbf{I} + \mathbf{K}_s & \mathbf{0} \\ \hline \mathbf{0} & \sigma_n^2\mathbf{I} \end{array}\right] \tag{422}$$

and

$$\mathbf{K}_0 = \left[\begin{array}{c|c} \sigma_n^2\mathbf{I} & \mathbf{0} \\ \hline \mathbf{0} & \sigma_n^2\mathbf{I} + \mathbf{K}_s \end{array}\right], \tag{423}$$

where we have partitioned the $2N \times 2N$ matrices into $N \times N$ submatrices. Then

$$\Delta\mathbf{Q} = \left[\begin{array}{c|c} \dfrac{1}{\sigma_n^2}\mathbf{I} & \mathbf{0} \\ \hline \mathbf{0} & (\sigma_n^2\mathbf{I} + \mathbf{K}_s)^{-1} \end{array}\right] - \left[\begin{array}{c|c} (\sigma_n^2\mathbf{I} + \mathbf{K}_s)^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \dfrac{1}{\sigma_n^2}\mathbf{I} \end{array}\right]. \tag{424}$$

Using (397), we have

$$\Delta\mathbf{Q} = \frac{1}{\sigma_n^2}\left[\begin{array}{c|c} \mathbf{H} & \mathbf{0} \\ \hline \mathbf{0} & -\mathbf{H} \end{array}\right], \tag{425}$$

where, as previously defined in (397), $\mathbf{H}$ is

$$\mathbf{H} \triangleq (\sigma_n^2\mathbf{I} + \mathbf{K}_s)^{-1}\mathbf{K}_s. \tag{426}$$

If we partition $\mathbf{R}$ into two $N \times 1$ matrices,

$$\mathbf{R} = \left[\begin{array}{c} \mathbf{R}_1 \\ \hline \mathbf{R}_2 \end{array}\right], \tag{427}$$

then

$$l(\mathbf{R}) = \frac{1}{\sigma_n^2}\mathbf{R}_1^T\mathbf{H}\mathbf{R}_1 - \mathbf{R}_2^T\mathbf{H}\mathbf{R}_2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'. \tag{428}$$

The special cases analogous to 1A and 1B follow easily.

Case 2A. Uncorrelated, Identically Distributed Signal Components. Let

$$\mathbf{K}_s = \sigma_s^2\mathbf{I}; \tag{429}$$

then

$$l(\mathbf{R}) = \sum_{i=1}^{N} R_i^2 - \sum_{i=N+1}^{2N} R_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma^v. \tag{430}$$

If the hypotheses are equally likely and the criterion is minimum Pr(ϵ), the threshold η in the LRT is unity (see 69). From (388) and (390) we see that this will result in $\gamma^v = 0$. This case occurs frequently and leads to a simple error calculation. The test then becomes

$$l_1(\mathbf{R}) \triangleq \sum_{i=1}^{N} R_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} \sum_{i=N+1}^{2N} R_i^2 \triangleq l_0(\mathbf{R}). \tag{431}$$

The probability of error given that H_1 is true is the probability that $l_0(\mathbf{R})$ is greater than $l_1(\mathbf{R})$. Because the test is symmetric with respect to the two hypotheses,

$$\Pr(\epsilon) = \tfrac{1}{2}\Pr(\epsilon|H_1) + \tfrac{1}{2}\Pr(\epsilon|H_0) = \Pr(\epsilon|H_1). \tag{432a}$$

Thus

$$\Pr(\epsilon) = \int_0^\infty dL_1 p_{l_1|H_1}(L_1|H_1) \int_{L_1}^\infty p_{l_0|H_1}(L_0|H_1)\, dL_0. \tag{432b}$$

Substituting (406) and (407) in (432b), recalling that N is even, and evaluating the inner integral, we have

$$\Pr(\epsilon) = \int_0^\infty \frac{1}{2^{N/2}\sigma_1{}^N\Gamma(N/2)} L_1^{N/2-1}\, e^{-L_1/2\sigma_1{}^2} \times \left[e^{-L_1/2\sigma_n{}^2} \sum_{k=0}^{N/2-1} \frac{(L_1/2\sigma_n{}^2)^k}{k!}\right] dL_1. \tag{432c}$$

Defining

$$\alpha = \frac{\sigma_n^2}{\sigma_1^2 + \sigma_n^2}, \tag{433}$$

and integrating, (432c) reduces to

$$\Pr(\epsilon) = \alpha^{N/2} \sum_{j=0}^{N/2-1} \binom{\frac{N}{2} + j - 1}{j} (1 - \alpha)^j. \tag{434}$$

This result is due to Pierce [22]. It is a closed-form expression but it is tedious to use. We delay plotting (434) until Section 2.7, in which we derive an approximate expression for comparison.

Case 2B. Uncorrelated Signal Components: Unequal Variances. Now,

$$\mathbf{K}_s = \begin{bmatrix} \sigma_{s_1}{}^2 & & & 0 \\ & \sigma_{s_2}{}^2 & & \\ & & \ddots & \\ 0 & & & \sigma_{s_N}{}^2 \end{bmatrix}. \tag{435}$$

It follows easily that

$$l(\mathbf{R}) = \frac{1}{\sigma_n^2}\left[\sum_{i=1}^{N} \frac{\sigma_{s_i}{}^2}{\sigma_n{}^2 + \sigma_{s_i}{}^2} R_i^2 - \sum_{i=N+1}^{2N} \frac{\sigma_{s_i-N}^2}{\sigma_n{}^2 + \sigma_{s_i-N}^2} R_i{}^2\right] \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \gamma'. \tag{436}$$

As in Case 1B, the performance is difficult to evaluate. The approximations developed in Section 2.7 are also useful for this case.

2.6.3 Summary

We have discussed in detail the general Gaussian problem and have found that the sufficient statistic was the difference between two quadratic forms:

$$l(\mathbf{R}) = \tfrac{1}{2}(\mathbf{R}^T - \mathbf{m}_0{}^T)\mathbf{Q}_0(\mathbf{R} - \mathbf{m}_0) - \tfrac{1}{2}(\mathbf{R}^T - \mathbf{m}_1{}^T)\mathbf{Q}_1(\mathbf{R} - \mathbf{m}_1). \tag{437}$$

A particularly simple special case was the one in which the covariance matrices on the two hypotheses were equal. Then

$$l(\mathbf{R}) = \tfrac{1}{2}\,\Delta\mathbf{m}^T\mathbf{Q}\mathbf{R}, \tag{438}$$

and the performance was completely characterized by the quantity d^2:

$$d^2 = \Delta\mathbf{m}^T\mathbf{Q}\,\Delta\mathbf{m}. \tag{439}$$

When the covariance matrices are unequal, the implementation of the likelihood ratio test is still straightforward but the performance calculations are difficult (remember that d^2 is no longer applicable because $l(\mathbf{R})$ is not Gaussian). In the simplest case of diagonal covariance matrices with equal elements, exact error expressions were developed. In the general case, exact expressions are possible but are too unwieldy to be useful. This inability to obtain tractable performance expressions is the motivation for discussion of performance bounds and approximations in the next section.

Before leaving the general Gaussian problem, we should point out that similar results can be obtained for the M-hypothesis case and for the estimation problem. Some of these results are developed in the problems.

2.7 PERFORMANCE BOUNDS AND APPROXIMATIONS

Up to this point we have dealt primarily with problems in which we could derive the structure of the optimum receiver and obtain relatively simple expressions for the receiver operating characteristic or the error probability.

In many cases of interest the optimum test can be derived but an exact performance calculation is impossible. For these cases we must resort to bounds on the error probabilities or approximate expressions for these probabilities. In this section we derive some simple bounds and approximations which are useful in many problems of practical importance. The basic results, due to Chernoff [28], were extended initially by Shannon [23]. They have been further extended by Fano [24], Shannon, Gallager, and Berlekamp [25], and Gallager [26] and applied to a problem of interest to us by Jacobs [27]. Our approach is based on the last two references. Because the latter part of the development is heuristic in nature, the interested reader should consult the references given for more careful derivations. From the standpoint of use in later sections, we shall not use the results until Chapter II-3 (the results are also needed for some of the problems in Chapter 4).

The problem of interest is the general binary hypothesis test outlined in Section 2.2. From our results in that section we know that it will reduce to a likelihood ratio test. We begin our discussion at this point.

The likelihood ratio test is

$$l(\mathbf{R}) \triangleq \ln \Lambda(\mathbf{R}) = \ln\left[\frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)}\right] \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{440}$$

The variable $l(\mathbf{R})$ is a random variable whose probability density depends on which hypothesis is true. In Fig. 2.36 we show a typical $p_{l|H_1}(L|H_1)$ and $p_{l|H_0}(L|H_0)$.

If the two densities are known, then P_F and P_D are given by

$$P_D = \int_{\gamma}^{\infty} p_{l|H_1}(L|H_1)\,dL, \tag{441}$$

$$P_F = \int_{\gamma}^{\infty} p_{l|H_0}(L|H_0)\,dL. \tag{442}$$

The difficulty is that it is often hard to find $p_{l|H_i}(L|H_i)$, and even if it can be found it is cumbersome. Typical of this complexity is Case 1A

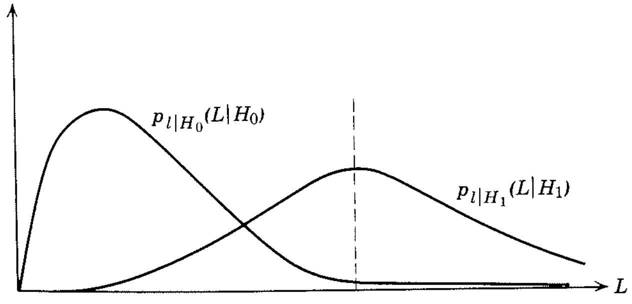


Fig. 2.36 Typical densities.

on p. 108, in which there are N Gaussian variables with equal variances making up the signal. To analyze a given system, the errors may be evaluated numerically. On the other hand, if we set out to synthesize a system, it is inefficient (if not impossible) to try successive systems and evaluate each numerically. Therefore we should like to find some simpler approximate expressions for the error probabilities.

In this section we derive some simple expressions that we shall use in the sequel. We first focus our attention on cases in which $l(\mathbf{R})$ is a sum of independent random variables. This suggests that its characteristic function may be useful, for it will be the product of the individual characteristic functions of the R_i. Similarly, the moment-generating function will be the product of individual moment-generating functions. Therefore an approximate expression based on one of these functions should be relatively easy to evaluate. The first part of our discussion develops bounds on the error probabilities in terms of the moment-generating function of $l(\mathbf{R})$.

In the second part we consider the case in which $l(\mathbf{R})$ is the sum of a *large* number of independent random variables. By the use of the central limit theorem we improve on the results obtained in the first part of the discussion.

We begin by deriving a simple upper bound on P_F in terms of the moment-generating function. The moment-generating function of $l(\mathbf{R})$ on hypothesis H_0 is

$$\phi_{l|H_0}(s) \triangleq E(e^{sl}|H_0) = \int_{-\infty}^{\infty} e^{sL} p_{l|H_0}(L|H_0)\, dL, \tag{443}$$

where s is a *real* variable. (The range of s corresponds to those values for which the integral exists.) We shall see shortly that it is more useful to write

$$\phi_{l|H_0}(s) \triangleq \exp\,[\mu(s)], \tag{444}$$

so that

$$\mu(s) = \ln \int_{-\infty}^{\infty} e^{sL} p_{l|H_0}(L|H_0)\, dL. \tag{445}$$

We may also express $\mu(s)$ in terms of $p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)$ and $p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)$. Because l is just a function of $\mathbf{r}$, we can write (443) as

$$\phi_{l|H_0}(s) = \int_{-\infty}^{\infty} e^{sl(\mathbf{R})} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R}. \tag{446}$$

Then

$$\mu(s) = \ln \int_{-\infty}^{\infty} e^{sl(\mathbf{R})} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R}. \tag{447}$$

Using (440),

$$\mu(s) = \ln \int_{-\infty}^{\infty} \left(\frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)}\right)^{s} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R}, \tag{448}$$

or

$$\boxed{\mu(s) = \ln \int_{-\infty}^{\infty} [p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)]^{s} [p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)]^{1-s}\, d\mathbf{R}.} \tag{449}$$

The function $\mu(s)$ plays a central role in the succeeding discussion. It is now convenient to rewrite the error expressions in terms of a new random variable whose mean is in the vicinity of the threshold. The reason for this step is that we shall use the central limit theorem in the second part of our derivation. It is most effective near the mean of the random variable of interest. Consider the simple probability density shown in Fig. 2.37*a*. To get the new family of densities shown in Figs. 2.37*b–d* we multiply $p_x(X)$ by e^{sX} for various values of s (and normalize to obtain a unit area). We see that for $s > 0$ the mean is shifted to the right. For the moment we leave s as a parameter. We see that increasing s "tilts" the density more.

Denoting this new variable as x_s, we have

$$p_{x_s}(X) \triangleq \frac{e^{sX} p_{l|H_0}(X|H_0)}{\int_{-\infty}^{\infty} e^{sL} p_{l|H_0}(L|H_0)\, dL} = \frac{e^{sX} p_{l|H_0}(X|H_0)}{e^{\mu(s)}}. \tag{450}$$

Observe that we define x_s in terms of its density function, for that is what we are interested in. Equation 450 is a general definition. For the density shown in Fig. 2.37, the limits would be $(-A, A)$.

We now find the mean and variance of x_s:

$$E(x_s) = \int_{-\infty}^{\infty} X p_{x_s}(X)\, dX = \frac{\int_{-\infty}^{\infty} X e^{sX} p_{l|H_0}(X|H_0)\, dX}{\int_{-\infty}^{\infty} e^{sL} p_{l|H_0}(L|H_0)\, dL}. \tag{451}$$

Comparing (451) and (445), we see that

$$E(x_s) = \frac{d\mu(s)}{ds} \triangleq \dot{\mu}(s). \tag{452}$$

Similarly, we find

$$\mathrm{Var}\,(x_s) = \ddot{\mu}(s). \tag{453}$$

[Observe that (453) implies that $\mu(s)$ is convex.]

We now rewrite P_F in terms of this tilted variable x_s:

$$P_F = \int_\gamma^\infty p_{l|H_0}(L|H_0)\,dL = \int_\gamma^\infty e^{\mu(s)-sX}p_{x_s}(X)\,dX$$
$$= e^{\mu(s)}\int_\gamma^\infty e^{-sX}p_{x_s}(X)\,dX. \tag{454}$$

We can now find a simple upper bound on P_F. For values of $s \geq 0$,

$$e^{-sX} \leq e^{-s\gamma}, \qquad \text{for } X \geq \gamma. \tag{455}$$

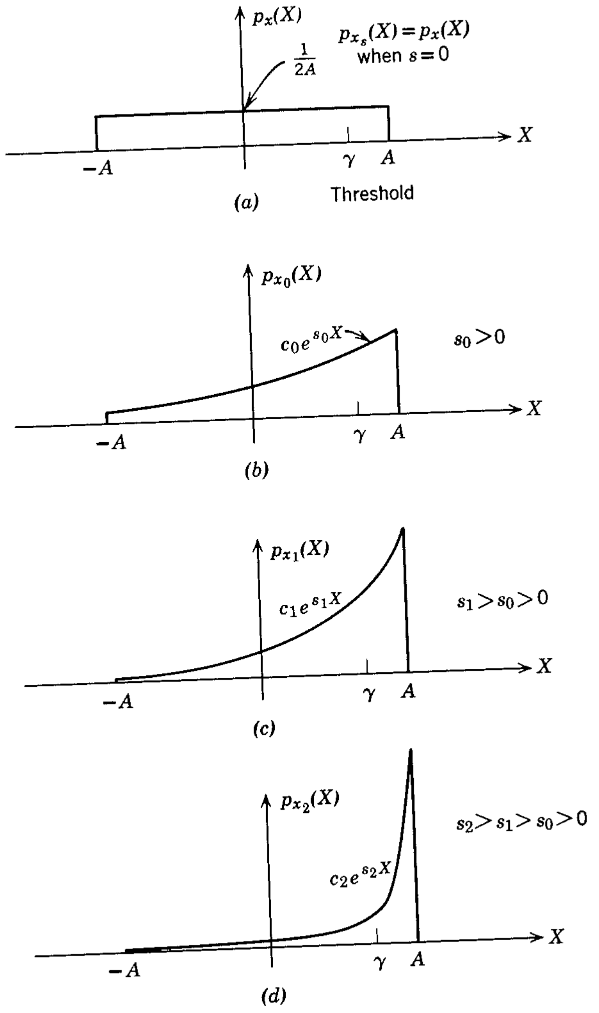


Fig. 2.37 Tilted probability densities.

Thus

$$P_F \leq e^{\mu(s)-s\gamma}\int_\gamma^\infty p_{x_s}(X)\,dX, \qquad s \geq 0. \tag{456}$$

Clearly the integral is less than one. Thus

$$P_F \leq e^{\mu(s)-s\gamma}, \qquad s \geq 0. \tag{457}$$

To get the best bound we minimize the right-hand side of (457) with respect to s. Differentiating the exponent and setting the result equal to zero, we obtain

$$\dot{\mu}(s) = \gamma. \tag{458}$$

Because $\ddot{\mu}(s)$ is nonnegative, a solution will exist if

$$\dot{\mu}(0) \leq \gamma \leq \dot{\mu}(\infty). \tag{459}$$

Because

$$\dot{\mu}(0) = E(l|H_0), \tag{460}$$

the left inequality implies that the threshold must be to the right of the mean of l on H_0. Assuming that (459) is valid, we have the desired result:

$$P_F \leq \exp\,[\mu(s) - s\dot{\mu}(s)], \qquad s \geq 0, \tag{461}$$

where s satisfies (458). (We have assumed $\mu(s)$ exists for the desired s.)

Equation 461 is commonly referred to as the Chernoff bound [28]. Observe that s is chosen so that the mean of the tilted variable x_s is at the threshold.

The next step is find a bound on P_M, the probability of a miss:

$$P_M = \int_{-\infty}^{\gamma} p_{l|H_1}(X|H_1)\,dX, \tag{462}$$

which we want to express in terms of the tilted variable x_s.

Using an argument identical to that in (88) through (94), we see that

$$p_{l|H_1}(X|H_1) = e^X p_{l|H_0}(X|H_0). \tag{463}$$

Substituting (463) into the right side of (450), we have

$$p_{l|H_1}(X|H_1) = e^{\mu(s)+(1-s)X}p_{x_s}(X). \tag{464}$$

Substituting into (462),

$$P_M = e^{\mu(s)}\int_{-\infty}^{\gamma} e^{(1-s)X}p_{x_s}(X)\,dX. \tag{465}$$

For $s \leq 1$

$$e^{(1-s)X} \leq e^{(1-s)\gamma}, \qquad \text{for } X \leq \gamma. \tag{466}$$

Thus

$$P_M \leq e^{\mu(s)+(1-s)\gamma}\int_{-\infty}^{\gamma} p_{x_s}(X)\,dX$$
$$\leq e^{\mu(s)+(1-s)\gamma}, \qquad s \leq 1. \tag{467}$$

Once again the bound is minimized for

$$\gamma = \dot{\mu}(s) \tag{468}$$

if a solution exists for $s \leq 1$. Observing that

$$\dot{\mu}(1) = E(l|H_1), \tag{469}$$

we see that this requires the threshold to be to the left of the mean of l on H_1.

Combining (461) and (467), we have

$$\boxed{\begin{aligned} P_F &\leq \exp\,[\mu(s) - s\dot{\mu}(s)] \\ & & 0 \leq s \leq 1 \\ P_M &\leq \exp\,[\mu(s) + (1 - s)\dot{\mu}(s)] \end{aligned}} \tag{470}$$

and

$$\gamma = \dot{\mu}(s)$$

is the threshold that lies *between* the means of l on the two hypotheses. Confining s to [0, 1] is not too restrictive because if the threshold is not between the means the error probability will be large on one hypothesis (greater than one half if the median coincides with the mean). If we are modeling some physical system this would usually correspond to unacceptable performance and necessitate a system change.

As pointed out in [25], the exponents have a simple graphical interpretation. A typical $\mu(s)$ is shown in Fig. 2.38. We draw a tangent at the point at which $\dot{\mu}(s) = \gamma$. This tangent intersects vertical lines at $s = 0$ and $s = 1$. The value of the intercept at $s = 0$ is the exponent in the P_F bound. The value of the intercept at $s = 1$ is the exponent in the P_M bound.

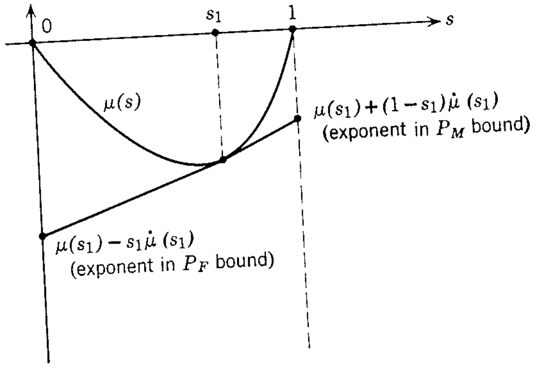

0
s_1
1
s
$\mu(s)$
$\mu(s_1)+(1-s_1)\dot{\mu}\,(s_1)$
(exponent in P_M bound)
$\mu(s_1)-s_1\dot{\mu}\,(s_1)$
(exponent in P_F bound)

Fig. 2.38 Exponents in bounds.

For the special case in which the hypotheses are equally likely and the error costs are equal we know that $\gamma = 0$. Therefore to minimize the bound we choose that value of s where $\dot{\mu}(s) = 0$.

The probability of error Pr (ϵ) is

$$\Pr(\epsilon) = \tfrac{1}{2}P_F + \tfrac{1}{2}P_M. \tag{471}$$

Substituting (456) and (467) into (471) and denoting the value s for which $\dot{\mu}(s) = 0$ as s_m, we have

$$\Pr(\epsilon) \leq \tfrac{1}{2}e^{\mu(s_m)} \int_0^\infty p_{x_s}(X)\,dX + \tfrac{1}{2}e^{\mu(s_m)} \int_{-\infty}^0 p_{x_s}(X)\,dX, \tag{472}$$

or

$$\boxed{\Pr(\epsilon) \leq \tfrac{1}{2}e^{\mu(s_m)}.} \tag{473}$$

Up to this point we have considered arbitrary binary hypothesis tests. The bounds in (470) and (473) are always valid if $\mu(s)$ exists. In many cases of interest $l(\mathbf{R})$ consists of a sum of a large number of independent random variables, and we can obtain a simple approximate expression for P_F and P_M that provides a much closer estimate of their actual value than the above bounds. The exponent in this expression is the same, but the multiplicative factor will often be appreciably smaller than unity.

We start the derivation with the expression for P_F given in (454). Motivated by our result in the bound derivation (458), we choose s so that

$$\dot{\mu}(s) = \gamma.$$

Then (454) becomes

$$P_F = e^{\mu(s)} \int_{\dot{\mu}(s)}^\infty e^{-sX} p_{x_s}(X)\,dX. \tag{474}$$

This can be written as

$$P_F = e^{\mu(s) - s\dot{\mu}(s)} \int_{\dot{\mu}(s)}^\infty e^{+s[\dot{\mu}(s) - X]} p_{x_s}(X)\,dX. \tag{475}$$

The term outside is just the bound in (461). We now use a central limit theorem argument to evaluate the integral. First define a standardized variable:

$$y \triangleq \frac{x_s - E(x_s)}{(\mathrm{Var}\,[x_s])^{1/2}} = \frac{x_s - \dot{\mu}(s)}{\sqrt{\ddot{\mu}(s)}}. \tag{476}$$

Substituting (476) into (475), we have

$$P_F = e^{\mu(s) - s\dot{\mu}(s)} \int_0^\infty e^{-s\sqrt{\ddot{\mu}(s)}\,Y} p_y(Y)\,dY. \tag{477}$$

In many cases the probability density governing $\mathbf{r}$ is such that y approaches a Gaussian random variable as N (the number of components of $\mathbf{r}$) approaches infinity.† A simple case in which this is true is the case in which the r_i are independent, identically distributed random variables with finite means and variances. In such cases, y approaches a zero-mean Gaussian random variable with unit variance and the integral in (477) can be evaluated by substituting the limiting density.

$$\int_0^\infty e^{-s\sqrt{\ddot{\mu}(s)}\,Y} \frac{1}{\sqrt{2\pi}} e^{-(Y^2/2)}\, dY = e^{s^2\ddot{\mu}(s)/2} \operatorname{erfc}_* (s\sqrt{\ddot{\mu}(s)}). \tag{478}$$

Then

$$P_F \simeq \left\{\exp\left[\mu(s) - s\dot{\mu}(s) + \frac{s^2}{2}\ddot{\mu}(s)\right]\right\} \operatorname{erfc}_* [s\sqrt{\ddot{\mu}(s)}]. \tag{479}$$

The approximation arises because y is only approximately Gaussian for finite N. For values of $s\sqrt{\ddot{\mu}(s)} > 3$ we can approximate $\operatorname{erfc}_*(\cdot)$ by the upper bound in (71). Using this approximation,

$$\boxed{P_F \simeq \frac{1}{\sqrt{2\pi s^2\ddot{\mu}(s)}} \exp\,[\mu(s) - s\dot{\mu}(s)], \qquad s \geq 0.} \tag{480}$$

It is easy to verify that the approximate expression in (480) can also be obtained by letting

$$p_y(Y) \simeq p_y(0) \simeq \frac{1}{\sqrt{2\pi}}. \tag{481}$$

Looking at Fig. 2.39, we see that this is valid when the exponential function decreases to a small value while $Y \ll 1$.

In exactly the same manner we obtain

$$P_M \simeq \left\{\exp\left[\mu(s) + (1 - s)\,\dot{\mu}(s) + \frac{(s-1)^2}{2}\ddot{\mu}(s)\right]\right\} \operatorname{erfc}_* [(1 - s)\sqrt{\ddot{\mu}(s)}]. \tag{482}$$

For $(1 - s)\sqrt{\ddot{\mu}(s)} > 3$, this reduces to

$$\boxed{P_M \simeq \frac{1}{\sqrt{2\pi(1-s)^2\ddot{\mu}(s)}} \exp\,[\mu(s) + (1 - s)\dot{\mu}(s)], \qquad s \leq 1.} \tag{483}$$

Observe that the exponents in (480) and (483) are identical to those obtained by using the Chernoff bound. The central limit theorem argument has provided a multiplicative factor that will be significant in many of the applications of interest to us.

† An excellent discussion is contained in Feller [33], pp. 517–520.

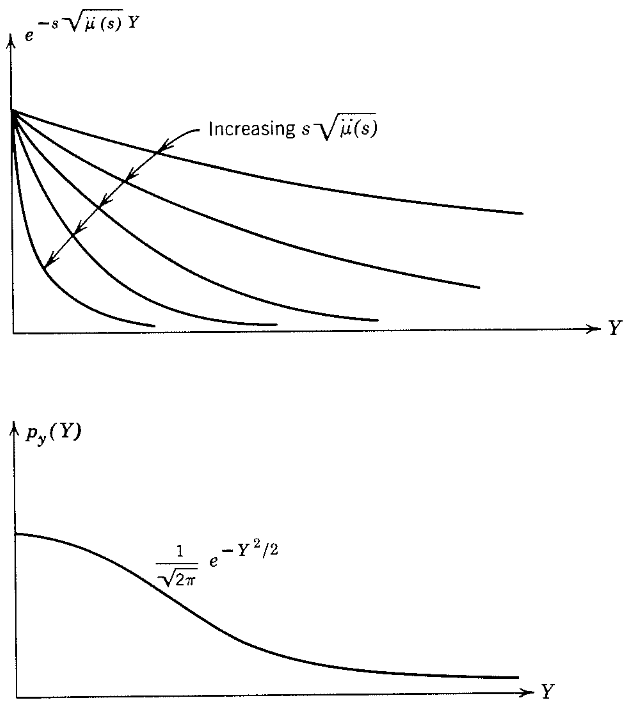


Fig. 2.39 Behavior of functions.

For the case in which $\Pr(\epsilon)$ is the criterion and the hypotheses are equally likely we have

$$\begin{aligned}\Pr(\epsilon) &= \tfrac{1}{2}P_F + \tfrac{1}{2}P_M \\ &= \tfrac{1}{2}\exp\left[\mu(s_m) + \frac{{s_m}^2}{2}\ddot{\mu}(s_m)\right] \operatorname{erfc}_* [s_m\sqrt{\ddot{\mu}(s_m)}] \\ &\quad + \tfrac{1}{2}\exp\left[\mu(s_m) + \frac{(1-s_m)^2}{2}\ddot{\mu}(s_m)\right] \operatorname{erfc}_* [(1 - s_m)\sqrt{\ddot{\mu}(s_m)}],\end{aligned} \tag{484}$$

where s_m is defined in the statement preceding (472) [i.e., $\dot{\mu}(s_m) = 0 = \gamma$]. When both $s_m\sqrt{\ddot{\mu}(s_m)} > 3$ and $(1 - s_m)\sqrt{\ddot{\mu}(s_m)} > 3$, this reduces to

$$\boxed{\Pr(\epsilon) \simeq \frac{1}{[2(2\pi\ddot{\mu}(s_m))^{1/2} s_m(1 - s_m)]} \exp \mu(s_m).} \tag{485}$$

We now consider several examples to illustrate the application of these ideas. The first is one in which the exact performance is known. We go through the bounds and approximations to illustrate the manipulations involved.

Example 1. In this example we consider the simple Gaussian problem first introduced on p. 27:

$$p_{\mathbf{r}|H_1}(\mathbf{R}|H_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(R_i - m)^2}{2\sigma^2}\right] \tag{486}$$

and

$$p_{\mathbf{r}|H_0}(\mathbf{R}|H_0) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R_i^2}{2\sigma^2}\right). \tag{487}$$

Then, using (449)

$$\mu(s) = \ln \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(R_i - m)^2 s + R_i^2(1 - s)}{2\sigma^2}\right] dR_i. \tag{488a}$$

Because all the integrals are identical,

$$\mu(s) = N \ln \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(R - m)^2 s + R^2(1 - s)}{2\sigma^2}\right] dR. \tag{488b}$$

Integrating we have

$$\mu(s) = Ns(s - 1)\frac{m^2}{2\sigma^2} \triangleq \frac{s(s - 1)d^2}{2}, \tag{489}$$

where d^2 was defined in the statement after (64). The curve is shown in Fig. 2.40:

$$\dot{\mu}(s) = \frac{(2s - 1)d^2}{2}. \tag{490}$$

Using the bounds in (470), we have

$$\left.\begin{aligned} P_F &\leq \exp\left(\frac{-s^2 d^2}{2}\right) \\ P_M &\leq \exp\left[-\frac{(1 - s)^2 d^2}{2}\right] \end{aligned}\right\} \quad 0 \leq s \leq 1. \tag{491}$$

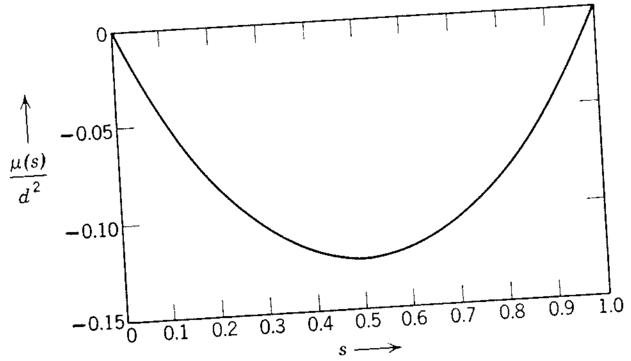


Fig. 2.40 $\mu(s)$ **for Gaussian variables with unequal means.**

Because $l(\mathbf{R})$ is the sum of Gaussian random variables, the expressions in (479) and (482) are exact. Evaluating $\ddot{\mu}(s)$, we obtain

$$\ddot{\mu}(s) = d^2. \tag{492}$$

Substituting into (479) and (482), we have

$$P_F = \text{erfc}_* \left[s\sqrt{\ddot{\mu}(s)}\right] = \text{erfc}_*(sd) \tag{493}$$

and

$$P_M = \text{erfc}_* \left[(1 - s)\sqrt{\ddot{\mu}(s)}\right] = \text{erfc}_* [(1 - s)d]. \tag{494}$$

These expressions are identical to (64) and (68) (let $s = (\ln \eta)/d^2 + \frac{1}{2}$).

An even simpler case is one in which the total probability of error is the criterion. Then we choose an s_m such as $\dot{\mu}(s_m) = 0$. From Fig. 2.40, we see that $s_m = \frac{1}{2}$. Using (484) and (485) we have

$$\Pr(\epsilon) = \text{erfc}_* \left(\frac{d}{2}\right) \simeq \left(\frac{2}{\pi d^2}\right)^{1/2} \exp\left(-\frac{d^2}{8}\right), \tag{495}$$

where the approximation is very good for $d > 6$.

This example is a special case of the binary symmetric hypothesis problem in which $\mu(s)$ is symmetric about $s = \frac{1}{2}$. When this is true *and* the criterion is minimum Pr (ϵ), then $\mu(\frac{1}{2})$ is the important quantity.

$$\mu(\tfrac{1}{2}) = \ln \int_{-\infty}^{\infty} [p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)]^{1/2} [p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)]^{1/2}\, d\mathbf{R}. \tag{496}$$

The negative of this quantity is frequently referred to as the Bhattacharyya distance (e.g., [29]). It is important to note that it is the significant quantity only when $s_m = \frac{1}{2}$.

In our next example we look at a more interesting case.

Example 2. This example is Case 1A of the general Gaussian problem described on p. 108:

$$\begin{aligned} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left(-\frac{R_i^2}{2\sigma_1^2}\right), \\ p_{\mathbf{r}|H_0}(\mathbf{R}|H_0) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left(-\frac{R_i^2}{2\sigma_0^2}\right). \end{aligned} \tag{497}$$

Substituting (497) into (499) gives,

$$\mu(s) = N \ln \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi}\,\sigma_1{}^s \sigma_0^{1-s})} \exp\left[-\frac{sR^2}{2\sigma_1^2} - \frac{(1 - s)R^2}{2\sigma_0^2}\right] dR \tag{498}$$

or

$$\mu(s) = \frac{N}{2} \ln \left[\frac{(\sigma_0^2)^s (\sigma_1^2)^{1-s}}{s\sigma_0^2 + (1 - s)\sigma_1^2}\right]. \tag{499}$$

A case that will be of interest in the sequel is

$$\begin{aligned} \sigma_1^2 &= \sigma_n^2 + \sigma_s^2, \\ \sigma_0^2 &= \sigma_n^2. \end{aligned} \tag{500}$$

Substituting (500) into (499) gives

$$\frac{\mu(s)}{N/2} = \left\{(1-s)\ln\left(1+\frac{\sigma_s^2}{\sigma_n^2}\right) - \ln\left[1+(1-s)\frac{\sigma_s^2}{\sigma_n^2}\right]\right\}. \tag{501}$$

This function is shown in Fig. 2.41.

$$\dot{\mu}(s) = \frac{N}{2}\left[-\ln\left(1+\frac{\sigma_s^2}{\sigma_n^2}\right) + \frac{\sigma_s^2/\sigma_n^2}{1+(1-s)\sigma_s^2/\sigma_n^2}\right] \tag{502}$$

and

$$\ddot{\mu}(s) = \frac{N}{2}\left[\frac{\sigma_s^2/\sigma_n^2}{1+(1-s)(\sigma_s^2/\sigma_n^2)}\right]^2. \tag{503}$$

By substituting (501), (502), and (503) into (479) and (482) we can plot an approximate receiver operating characteristic. This can be compared with the exact ROC in Fig. 2.35*a* to estimate the accuracy of the approximation. In Fig. 2.42 we show the comparison for $N = 4$ and 8, and $\sigma_s^2/\sigma_n^2 = 1$. The lines connect the equal threshold points. We see that the approximation is good. For larger N the exact and approximate ROC are identical for all practical purposes.

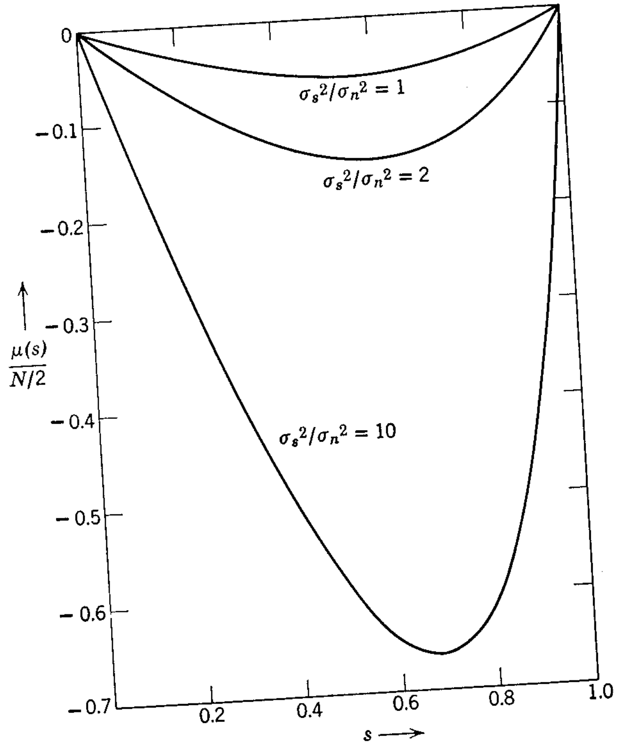


Fig. 2.41 $\mu(s)$ **for Gaussian variables with unequal variances.**

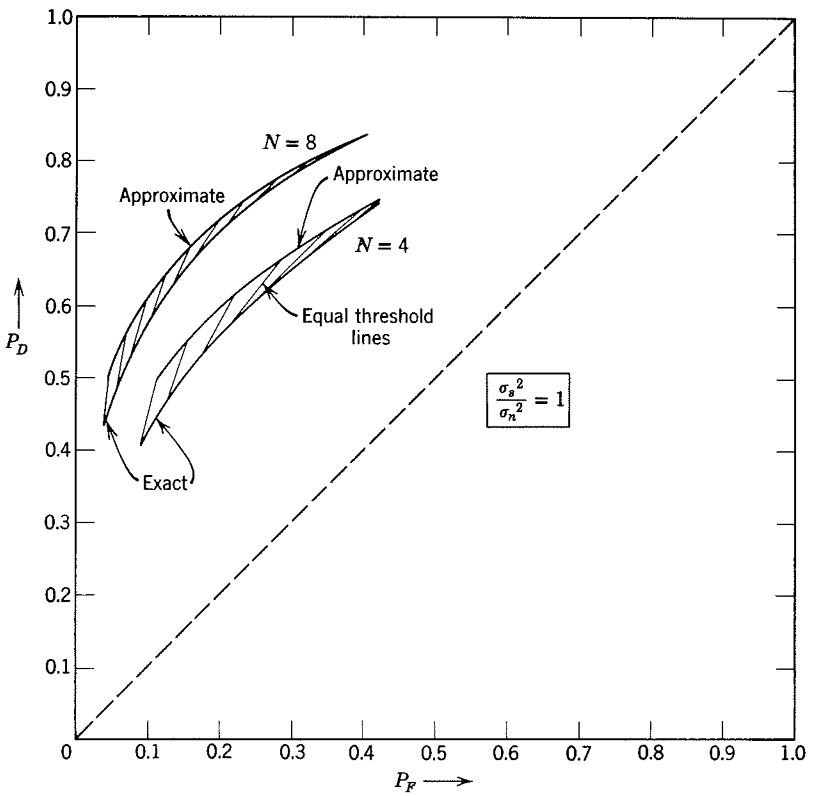


Fig. 2.42 Approximate receiver operating characteristics.

Example 3. In this example we consider first the simplified version of the symmetric hypothesis situation described in Case 2A (p. 115) in which $N = 2$.

$$p_{\mathbf{r}|H_1}(\mathbf{R}|H_1) = \frac{1}{(2\pi)^2\sigma_1^2\sigma_0^2}\exp\left(-\frac{R_1^2+R_2^2}{2\sigma_1^2} - \frac{R_3^2+R_4^2}{2\sigma_0^2}\right) \tag{504}$$

and

$$p_{\mathbf{r}|H_0}(\mathbf{R}|H_0) = \frac{1}{(2\pi)^2\sigma_1^2\sigma_0^2}\exp\left(-\frac{R_1^2+R_2^2}{2\sigma_0^2} - \frac{R_3^2+R_4^2}{2\sigma_1^2}\right), \tag{505}$$

where

$$\begin{aligned}\sigma_1^2 &= \sigma_s^2 + \sigma_n^2\\ \sigma_0^2 &= \sigma_n^2.\end{aligned} \tag{506}$$

Then

$$\begin{aligned}\mu(s) = {} & s\ln\sigma_n^2 + (1-s)\ln(\sigma_n^2+\sigma_s^2) - \ln(\sigma_n^2+\sigma_s^2 s)\\ & + (1-s)\ln\sigma_n^2 + s\ln(\sigma_n^2+\sigma_s^2) - \ln[\sigma_n^2+\sigma_s^2(1-s)]\\ = {} & \ln\left(1+\frac{\sigma_s^2}{\sigma_n^2}\right) - \ln\left[\left(1+\frac{s\sigma_s^2}{\sigma_n^2}\right)\left(1+\frac{(1-s)\sigma_s^2}{\sigma_n^2}\right)\right].\end{aligned} \tag{507}$$

The function $\mu(s)$ is plotted in Fig. 2.43*a*. The minimum is at $s = \frac{1}{2}$. This is the point of interest at which minimum Pr (ϵ) is the criterion.

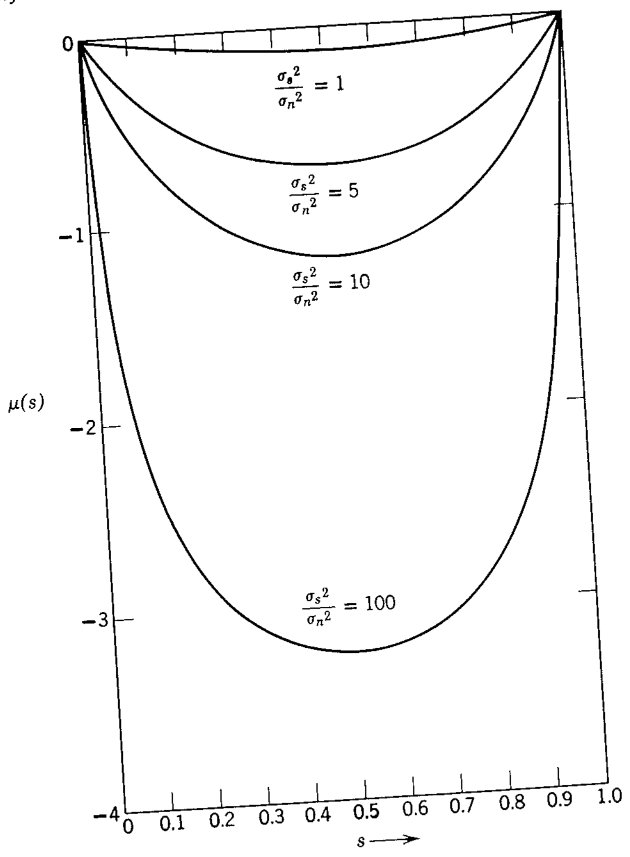


Fig. 2.43*a* **$\mu(s)$ for the binary symmetric hypothesis problem.**

Thus from (473), a *bound* on the error is,

$$\Pr(\epsilon) \le \tfrac{1}{2}\,\frac{(1 + \sigma_s^2/\sigma_n^2)}{(1 + \sigma_s^2/2\sigma_n^2)^2}. \tag{508}$$

The bound in (508) is plotted in Fig. 2.43*b*.

Example 3A. An interesting extension of Example 3 is the problem in which

$$\mathbf{K}_s = \begin{bmatrix} \sigma_1^2 & & & & & & & \\ & \sigma_1^2 & & & & 0 & & \\ & & \sigma_2^2 & & & & & \\ & & & \sigma_2^2 & & & & \\ & & & & \sigma_3^2 & & & \\ & & & & & \ddots & & \\ & 0 & & & & & \sigma_{N/2}^2 & \\ & & & & & & & \sigma_{N/2}^2 \end{bmatrix}. \tag{509}$$

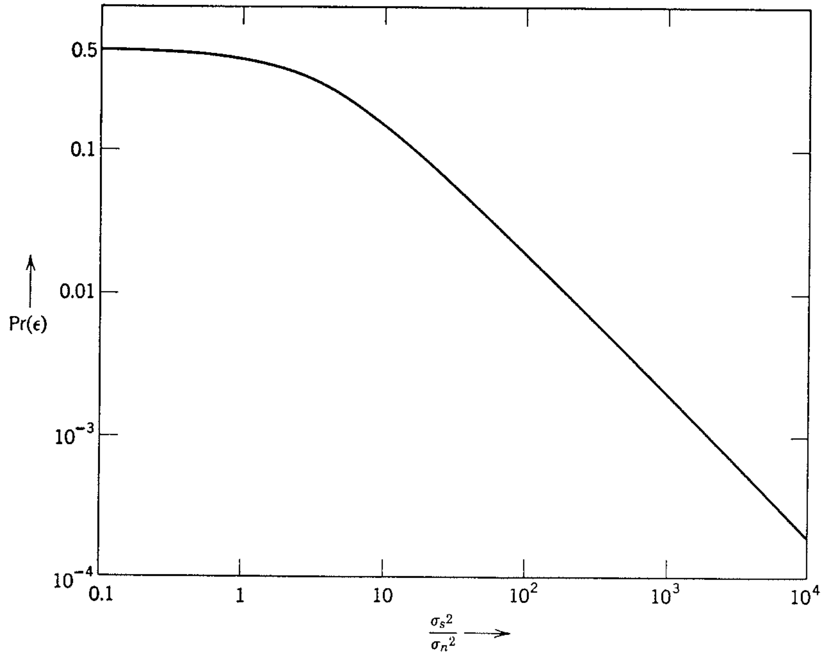


Fig. 2.43*b* **Bound on the probability of error ($\Pr(\epsilon)$).**

The r_i's are independent variables and their variances are pairwise equal. This is a special version of Case 2B on p. 115. We shall find later that it corresponds to a physical problem of appreciable interest.

Because of the independence, $\mu(s)$ is just the sum of the $\mu(s)$ for each pair, but each pair corresponds to the problem in Example 3. Therefore

$$\mu(s) = \sum_{i=1}^{N/2} \ln\left(1 + \frac{\sigma_{s_i}^2}{\sigma_n^2}\right) - \sum_{i=1}^{N/2} \ln\left\{\left(1 + s\,\frac{\sigma_{s_i}^2}{\sigma_n^2}\right)\left(1 + (1-s)\,\frac{\sigma_{s_i}^2}{\sigma_n^2}\right)\right\}. \tag{510}$$

Then

$$\dot{\mu}(s) = -\sum_{i=1}^{N/2} \left[\frac{\sigma_{s_i}^2}{\sigma_n^2 + s\sigma_{s_i}^2} - \frac{\sigma_{s_i}^2}{\sigma_n^2 + (1-s)\sigma_{s_i}^2}\right] \tag{511}$$

and

$$\ddot{\mu}(s) = \sum_{i=1}^{N/2} \left\{\frac{\sigma_{s_i}^4}{(\sigma_n^2 + s\sigma_{s_i}^2)^2} + \frac{\sigma_{s_i}^4}{[\sigma_n^2 + (1-s)\sigma_{s_i}^2]^2}\right\}. \tag{512}$$

For a minimum probability of error criterion it is obvious from (511) that $s_m = \frac{1}{2}$. Using (485), we have

$$\Pr(\epsilon) \simeq \left[\pi \sum_{i=1}^{N/2} \frac{\sigma_{s_i}^4}{(\sigma_n^2 + \frac{1}{2}\sigma_{s_i}^2)^2}\right]^{-1/2} \exp\left[\sum_{i=1}^{N/2} \ln\left(1 + \frac{\sigma_{s_i}^2}{\sigma_n^2}\right) - 2\sum_{i=1}^{N/2} \ln\left(1 + \frac{\sigma_{s_i}^2}{2\sigma_n^2}\right)\right] \tag{513}$$

or

$$\Pr(\epsilon) \simeq \left[\pi \sum_{i=1}^{N/2} \frac{\sigma_{s_i}^4}{(\sigma_n^2 + \frac{1}{2}\sigma_{s_i}^2)^2}\right]^{-1/2} \prod_{i=1}^{N/2} \frac{\left(1 + \frac{\sigma_{s_i}^2}{\sigma_n^2}\right)}{\left(1 + \frac{\sigma_{s_i}^2}{2\sigma_n^2}\right)^2}. \tag{514}$$

For the special case in which the variances are equal

$$\sigma_{s_i}^2 = \sigma_s^2 \tag{515}$$

and (514) reduces to

$$\Pr(\epsilon) \simeq \sqrt{\frac{2}{\pi N}}\, \frac{(1 + \sigma_s^2/\sigma_n^2)^{N/2}}{(\sigma_s^2/\sigma_n^2)(1 + \sigma_s^2/2\sigma_n^2)^{N-1}}. \tag{516}$$

Alternately, we can use the approximation given by (484). For this case it reduces to

$$\Pr(\epsilon) \simeq \left[\frac{1 + \sigma_s^2/\sigma_n^2}{(1 + \sigma_s^2/2\sigma_n^2)^2}\right]^{N/2} \exp\left[\frac{N}{8}\left(\frac{\sigma_s^2/\sigma_n^2}{1 + \sigma_s^2/2\sigma_n^2}\right)^2\right] \operatorname{erfc}_*\left[\left(\frac{N}{4}\right)^{1/2}\left(\frac{\sigma_s^2/\sigma_n^2}{1 + \sigma_s^2/2\sigma_n^2}\right)\right]. \tag{517}$$

In Fig. 2.44 we have plotted the approximate Pr (ε) using (517) and exact Pr (ε) which was given by (434). We see that the approximation is excellent.

The principal results of this section were the bounds on P_F and P_M given in (470) and (473) and the approximate error expressions given in (479), (480), (482), (483), (484), and (485). These expressions will enable us to find performance results for a number of cases of physical interest.

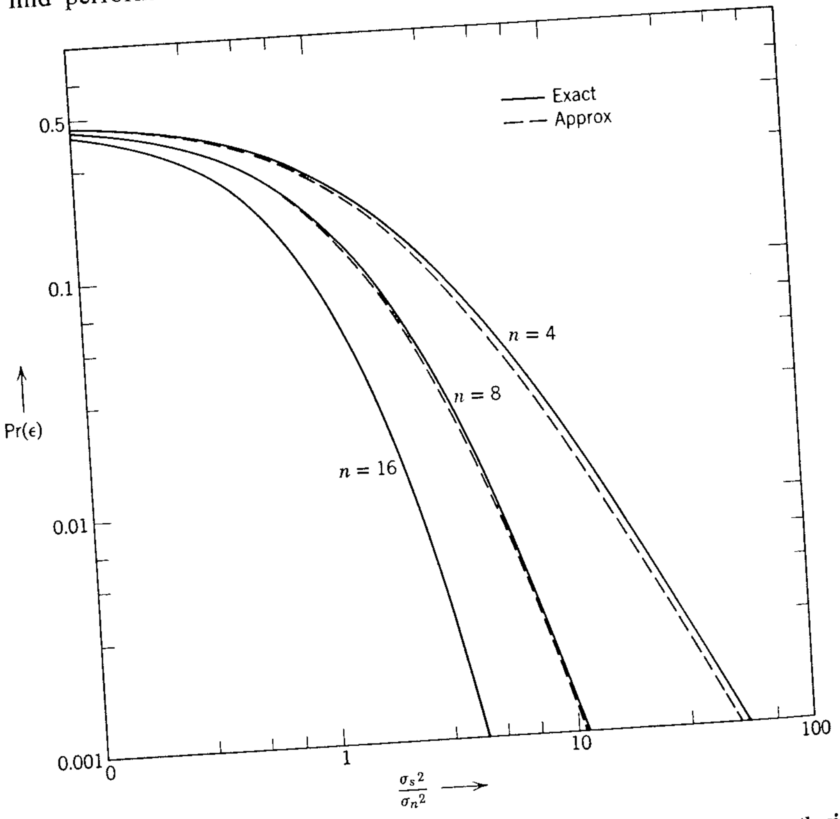


Fig. 2.44 Exact and approximate error expressions for the binary symmetric hypothesis case.

Results for some other cases are given in Yudkin [34] and Goblick [35] and the problems. In Chapter II-3 we shall study the detection of Gaussian signals in Gaussian noise. Suitable extensions of the above bounds and approximations will be used to evaluate the performance of the optimum processors.

2.8 SUMMARY

In this chapter we have derived the essential detection and estimation theory results that provide the basis for our work in the remainder of the book.

We began our discussion by considering the simple binary hypothesis testing problem. Using either a Bayes or a Neyman-Pearson criterion, we were led to a likelihood ratio test, whose performance was described by a receiver operating characteristic. Similarly, the M-hypothesis problem led to the construction of a set of likelihood ratios. This criterion-invariant reduction of the observation to a single number in the binary case or to $M - 1$ numbers in the M hypothesis case is the key to our ability to solve the detection problem when the observation is a waveform.

The development of the necessary estimation theory results followed a parallel path. Here, the fundamental quantity was a likelihood function. As we pointed out in Section 2.4, its construction is closely related to the construction of the likelihood ratio, a similarity that will enable us to solve many parallel problems by inspection. The composite hypothesis testing problem showed further how the two problems were related.

Our discussion through Section 2.5 was deliberately kept at a general level and for that reason forms a broad background of results applicable to many areas in addition to those emphasized in the remainder of the book. In Section 2.6 we directed our attention to the general Gaussian problem, a restriction that enabled us to obtain more specific results than were available in the general case. The waveform analog to this general Gaussian problem plays the central role in most of the succeeding work.

The results in the general Gaussian problem illustrated that although we can always find the optimum processor the exact performance may be difficult to calculate. This difficulty motivated our discussion of error bounds and approximations in Section 2.7. These approximations will lead us to useful results in several problem areas of practical importance.

2.9 PROBLEMS

The problems are divided into sections corresponding to the major sections in the chapter. For example, section P2.2 pertains to text material in Section 2.2. In sections in which it is appropriate the problems are divided into topical groups.

As pointed out in the Preface, solutions to individual problems are available on request.

P2.2 Binary Hypothesis Tests

Simple Binary Tests

Problem 2.2.1. Consider the following binary hypothesis testing problem:

$$H_1: r = s + n,$$
$$H_0: r = n,$$

where s and n are independent random variables.

$$p_s(S) = \begin{cases} ae^{-aS} & S \geq 0, \\ 0 & S < 0, \end{cases}$$

$$p_n(N) = \begin{cases} be^{-bN} & N \geq 0, \\ 0 & N < 0. \end{cases}$$

1. Prove that the likelihood ratio test reduces to

$$R \underset{H_0}{\overset{H_1}{\gtrless}} \gamma.$$

2. Find γ for the optimum Bayes test as a function of the costs and a priori probabilities.

3. Now assume that we need a Neyman-Pearson test. Find γ as a function of P_F, where

$$P_F \triangleq \Pr(\text{say } H_1 | H_0 \text{ is true}).$$

Problem 2.2.2. The two hypotheses are

$$H_1: p_r(R) = \frac{1}{2} \exp(-|R|)$$

$$H_0: p_r(R) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} R^2\right)$$

1. Find the likelihood ratio $\Lambda(R)$.
2. The test is

$$\Lambda(R) \underset{H_0}{\overset{H_1}{\gtrless}} \eta.$$

Compute the decision regions for various values of η.

Problem 2.2.3. The random variable x is $N(0, \sigma)$. It is passed through one of two nonlinear transformations.

$$H_1: y = x^2,$$
$$H_0: y = x^3.$$

Find the LRT.

Problem 2.2.4. The random variable x is $N(m, \sigma)$. It is passed through one of two nonlinear transformations.

$$H_1: y = e^x,$$
$$H_0: y = x^2.$$

Find the LRT.

Problem 2.2.5. Consider the following hypothesis-testing problem. There are K independent observations.

$$H_1: r_i \text{ is Gaussian, } N(0, \sigma_1), \qquad i = 1, 2, \ldots, K,$$
$$H_0: r_i \text{ is Gaussian, } N(0, \sigma_0), \qquad i = 1, 2, \ldots, K,$$

where $\sigma_0 < \sigma_1$.

1. Compute the likelihood ratio.
2. Assume that the threshold is η:

$$\Lambda(\mathbf{R}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta.$$

Show that a sufficient statistic is $l(\mathbf{R}) = \sum_{i=1}^{K} R_i^2$. Compute the threshold γ for the test

$$l(\mathbf{R}) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

in terms of η, σ_0, σ_1.

3. Define

$$P_F = \Pr(\text{choose } H_1 | \, H_0 \text{ is true}),$$
$$P_M = \Pr(\text{choose } H_0 | \, H_1 \text{ is true}).$$

Find an expression for P_F and P_M.

4. Plot the ROC for $K = 1$, $\sigma_1^2 = 2$, $\sigma_0^2 = 1$.
5. What is the threshold for the minimax criterion when $C_M = C_F$ and $C_{00} = C_{11} = 0$?

Problem 2.2.6. The observation r is defined in the following manner:

$$r = bm_1 + n : H_1,$$
$$r = n \qquad : H_0,$$

where b and n are independent zero-mean Gaussian variables with variances σ_b^2 and σ_n^2, respectively

1. Find the LRT and draw a block diagram of the optimum processor.
2. Draw the ROC.
3. Assume that the two hypotheses are equally likely. Use the criterion of minimum probability of error. What is the $\Pr(\epsilon)$?

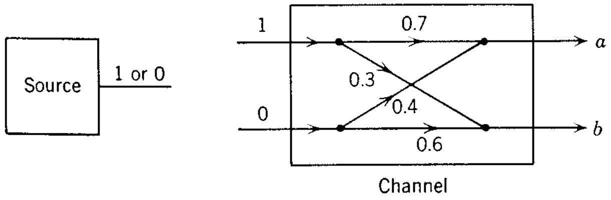
Problem 2.2.7. One of two possible sources supplies the inputs to the simple communication channel as shown in the figure.

Both sources put out either 1 or 0. The numbers on the line are the channel transition probabilities; that is,

$$\Pr(a \text{ out} \mid 1 \text{ in}) = 0.7.$$

The source characteristics are

Source 1:	Pr(1) = 0.5,	Pr(0) = 0.5;
Source 2:	Pr(1) = 0.6,	Pr(0) = 0.4.



To put the problem in familiar notation, define

(a) false alarm—say source 2 when source 1 is present;
(b) detection—say source 2 when source 2 is present.

1. Compute the ROC of a test that maximizes P_D subject to the constraint that $P_F = \alpha$.
2. Describe the test procedure in detail for $\alpha = 0.25$.

Problem 2.2.8. The probability densities on the two hypotheses are

$$p_{x|H_i}(X|H_i) = \frac{1}{\pi[1 + (X - a_i)^2]} \qquad -\infty < X < \infty : H_i, \qquad i = 0, 1.$$

where $a_0 = 0$ and $a_1 = 1$.

1. Find the LRT.
2. Plot the ROC.

Problem 2.2.9. Consider a simple coin tossing problem:

$$\begin{aligned} H_1&: \text{ heads are up}, & \Pr[H_1] &\triangleq P_1, \\ H_0&: \text{ tails are up}, & \Pr[H_0] &\triangleq P_0 < P_1. \end{aligned}$$

N independent tosses of the coin are made. Show that the number of observed heads, N_H, is a sufficient statistic for making a decision between the two hypotheses.

Problem 2.2.10. A sample function of a simple Poisson counting process $N(t)$ is observed over the interval T:

$$\begin{aligned} &\text{hypothesis } H_1: && \text{the mean rate is } k_1 : \Pr(H_1) = \tfrac{1}{2}, \\ &\text{hypothesis } H_0: && \text{the mean rate is } k_0 : \Pr(H_0) = \tfrac{1}{2}. \end{aligned}$$

1. Prove that the number of events in the interval T is a "sufficient statistic" to choose hypothesis H_0 or H_1.
2. Assuming equal costs for the possible errors, derive the appropriate likelihood ratio test and the threshold.
3. Find an expression for the probability of error.

Problem 2.2.11. Let

$$y = \sum_{i=0}^{n} x_i,$$

where the x_i are statistically independent random variables with a Gaussian density $N(0, \sigma)$. The number of variables in the sum is a random variable with a Poisson distribution:

$$\Pr(n = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, \ldots.$$

We want to decide between the two hypotheses,

$$\begin{aligned} H_1 &: n \leq 1, \\ H_0 &: n > 1. \end{aligned}$$

Write an expression for the LRT.

Problem 2.2.12. Randomized tests. Our basic model of the decision problem in the text (p. 24) did not permit randomized decision rules. We can incorporate them by assuming that at each point $\mathbf{R}$ in Z we say H_1 with probability $\phi(\mathbf{R})$ and say H_0 with probability $1 - \phi(\mathbf{R})$. The model in the text is equivalent to setting $\phi(\mathbf{R}) = 1$ for all $\mathbf{R}$ in Z_1 and $\phi(\mathbf{R}) = 0$ for all $\mathbf{R}$ in Z_0.

1. We consider the Bayes criterion first. Write the risk for the above decision model.
2. Prove that a LRT minimizes the risk and a randomized test is *never* necessary.

3. Prove that the risk is constant over the interior of any straight-line segment on an ROC. Because straight-line segments are generated by randomized tests, this is an alternate proof of the result in Part 2.
4. Consider the Neyman-Pearson criterion. Prove that the optimum test always consists of either

(i) an ordinary LRT with $P_F = \alpha$ or
(ii) a probabilistic mixture of *two* ordinary likelihood ratio tests constructed as follows: Test 1: $\Lambda(\mathbf{R}) \overset{H_1}{\geq} \eta$ gives $P_F = \alpha^+$. Test 2: $\Lambda(\mathbf{R}) \overset{H_1}{>} \eta$ gives $P_F = \alpha^-$, where $[\alpha^-, \alpha^+]$ is the smallest interval containing α. $\phi(\mathbf{R})$ is 0 or 1 except for those $\mathbf{R}$ where $\phi(\mathbf{R}) = \eta$. (Find $\phi(\mathbf{R})$ for this set.)

Mathematical Properties

Problem 2.2.13. The random variable $\Lambda(\mathbf{R})$ is defined by (13) and has a different probability density on H_1 and H_0. Prove the following:

1. $E(\Lambda^n|H_1) = E(\Lambda^{n+1}|H_0)$,
2. $E(\Lambda|H_0) = 1$,
3. $E(\Lambda|H_1) - E(\Lambda|H_0) = \text{Var}(\Lambda|H_0)$.

Problem 2.2.14. Consider the random variable Λ. In (94) we showed that

$$p_{\Lambda|H_1}(X|H_1) = X p_{\Lambda|H_0}(X|H_0).$$

1. Verify this relation by direct calculation of $p_{\Lambda|H_1}(\cdot)$ and $p_{\Lambda|H_0}(\cdot)$ for the densities in Example 1 [p. 27, (19) and (20)].
2. On page 37 we saw that the performance of the test in Example 1 was completely characterized by d^2. Show that

$$d^2 = \ln[1 + \text{Var}(\Lambda|H_0)].$$

Problem 2.2.15. The function $\text{erfc}_*(X)$ is defined in (66):

1. Integrate by parts to establish the bound

$$\frac{1}{\sqrt{2\pi}\,X}\left(1 - \frac{1}{X^2}\right)\exp\left(-\frac{X^2}{2}\right) < \text{erfc}_*(X) < \frac{1}{\sqrt{2\pi}\,X}\exp\left(-\frac{X^2}{2}\right), \qquad X > 0.$$

2. Generalize part 1 to obtain the asymptotic series

$$\text{erfc}_*(X) = \frac{1}{\sqrt{2\pi}\,X} e^{-X^2/2}\left[1 + \sum_{m=1}^{n} (-1)^m \frac{1\cdot 3\cdots(2m-1)}{X^{2m}} + R_n\right].$$

The remainder is less than the magnitude of the $n + 1$ term and is the same sign. *Hint*. Show that the remainder is

$$R_n = \left[(-1)^n \frac{1\cdot 3\cdots(2n-1)}{X^{2n}}\right]\theta,$$

where

$$\theta = \int_0^\infty e^{-t}\left(1 + \frac{2t}{X}\right)^{-n-\frac{1}{2}} dt \leq 1.$$

3. Assume that $X = 3$. Calculate a simple bound on the *percentage* error when $\text{erfc}_*(3)$ is approximated by the first n terms in the asymptotic series. Evaluate this percentage error for $n = 2, 3, 4$ and compare the results. Repeat for $X = 5$.

Problem 2.2.16.

1. Prove

$$\text{erfc}_*(X) < \frac{1}{2}\exp\left(-\frac{X^2}{2}\right), \qquad X > 0.$$

Hint. Show

$$[\text{erfc}_*(X)]^2 = \Pr(x \geq X, y \geq X) < \Pr(x^2 + y^2 \geq 2X^2),$$

where x and y are independent zero-mean Gaussian variables with unit variance.

2. For what values of X is this bound better than (71)?

Higher Dimensional Decision Regions

A simple binary test can always be reduced to a one-dimensional decision region. In many cases the results are easier to interpret in two or three dimensions. Some typical examples are illustrated in this section.

Problem 2.2.17.

$$H_1: p_{x_1,x_2|H_1}(X_1, X_2|H_1) = \frac{1}{4\pi\sigma_1\sigma_0}\left[\exp\left(-\frac{X_1^2}{2\sigma_1^2} - \frac{X_2^2}{2\sigma_0^2}\right) + \exp\left(-\frac{X_1^2}{2\sigma_0^2} - \frac{X_2^2}{2\sigma_1^2}\right)\right], \quad -\infty < X_1, X_2 < \infty.$$

$$H_0: p_{x_1,x_2|H_0}(X_1, X_2|H_0) = \frac{1}{2\pi\sigma_0^2}\exp\left(-\frac{X_1^2}{2\sigma_0^2} - \frac{X_2^2}{2\sigma_0^2}\right), \qquad -\infty < X_1, X_2 < \infty.$$

1. Find the LRT.
2. Write an exact expression for P_D and P_F. Upper and lower bound P_D and P_F by modifying the region of integration in the exact expression.

Problem 2.2.18. The joint probability density of the random variables x_i ($i = 1, 2, .., M$) on H_1 and H_0 is

$$p_{\mathbf{x}|H_1}(\mathbf{X}|H_1) = \sum_{k=1}^{M} p_k \frac{1}{(2\pi\sigma^2)^{M/2}}\exp\left[-\frac{(X_k - m)^2}{2\sigma^2}\right]\prod_{i \neq k}^{M}\exp\left(-\frac{X_i^2}{2\sigma^2}\right),$$

where

$$\sum_{k=1}^{M} p_k = 1,$$

$$p_{\mathbf{x}|H_0}(\mathbf{X}|H_0) = \prod_{i=1}^{M}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{X_i^2}{2\sigma^2}\right) \qquad -\infty < X_i < \infty.$$

1. Find the LRT.
2. Draw the decision regions for various values of η in the X_1, X_2-plane for the special case in which $M = 2$ and $p_1 = p_2 = \frac{1}{2}$.
3. Find an upper and lower bound to P_F and P_D by modifying the regions of integration.

Problem 2.2.19. The probability density of r_i on the two hypotheses is

$$p_{r_i|H_k}(R_i|H_k) = \frac{1}{\sqrt{2\pi}\,\sigma_k}\exp\left[-\frac{(R_i - m_k)^2}{2\sigma_k^2}\right], \qquad i = 1, 2, \ldots, N, \quad k = 0, 1.$$

The observations are independent.

1. Find the LRT. Express the test in terms of the following quantities:

$$l_\alpha = \sum_{i=1}^{N} R_i,$$

$$l_\beta = \sum_{i=1}^{N} R_i^2.$$

2. Draw the decision regions in the l_α,l_β-plane for the case in which

$$2m_0 = m_1 > 0,$$
$$2\sigma_1 = \sigma_0.$$

Problem 2.2.20 (continuation).

1. Consider the special case

$$m_0 = 0,$$
$$\sigma_0 = \sigma_1.$$

Draw the decision regions and compute the ROC.

2. Consider the special case

$$m_0 = m_1 = 0,$$
$$\sigma_1^2 = \sigma_s^2 + \sigma_n^2,$$
$$\sigma_0 = \sigma_n.$$

Draw the decision regions.

Problem 2.2.21. A shell is fired at one of two targets: under H_1 the point of aim has coordinates x_1, y_1, z_1; under H_0 it has coordinates x_0, y_0, z_0. The distance of the actual landing point from the point of aim is a zero-mean Gaussian variable, $N(0, \sigma)$, in each coordinate. The variables are independent. We wish to observe the point of impact and guess which hypothesis is true.

1. Formulate this as a hypothesis-testing problem and compute the likelihood ratio. What is the simplest sufficient statistic? Is the ROC in Fig. 2.9*a* applicable? If so, what value of d^2 do we use?

2. Now include the effect of time. Under H_k the desired explosion time is t_k ($k = 1, 2$). The distribution of the actual explosion time is

$$p_{\tau|H_k}(\tau) = \frac{1}{\sqrt{2\pi}\,\sigma_t}\exp\left(-\frac{(\tau - t_k)^2}{2\sigma_t^2}\right), \qquad -\infty < \tau < \infty, \quad k = 1, 2.$$

Find the LRT and compute the ROC.

P2.3 M-Hypothesis Tests

Problem 2.3.1.

1. Verify that the M-hypothesis Bayes test always leads to a decision space whose dimension is less than or equal to $M - 1$.
2. Assume that the coordinates of the decision space are

$$\Lambda_k(\mathbf{R}) \triangleq \frac{p_{\mathbf{r}|H_k}(\mathbf{R}|H_k)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)}, \qquad k = 1, 2, \ldots, M - 1.$$

Verify that the decision boundaries are hyperplanes.

Problem 2.3.2. The formulation of the M-hypothesis problem in the text leads to an efficient decision space but loses some of the symmetry.

1. Starting with (98) prove that an equivalent form of the Bayes test is the following:

Compute

$$\beta_i \triangleq \sum_{j=0}^{M-1} C_{ij}\Pr(H_j|\mathbf{R}), \qquad i = 0, 1, \ldots, M - 1,$$

and choose the *smallest*.

2. Consider the special cost assignment

$$\begin{aligned} C_{ii} &= 0, \qquad i = 0, 1, 2, \ldots, M-1, \\ C_{ij} &= C, \qquad i \neq j, i, j = 0, 1, 2, \ldots, M-1. \end{aligned}$$

Show that an equivalent test is the following:

Compute

$$\Pr(H_i|\mathbf{R}), \qquad i = 0, 1, 2, \ldots, M-1,$$

and choose the *largest*.

Problem 2.3.3. The observed random variable is Gaussian on each of five hypotheses.

$$p_{r|H_k}(R|H_k) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(R-m_k)^2}{2\sigma^2}\right), \qquad \begin{aligned} &-\infty < R < \infty, \\ &k = 1, 2, \ldots, 5, \end{aligned}$$

where

$$\begin{aligned} m_1 &= -2m, \\ m_2 &= -m, \\ m_3 &= 0, \\ m_4 &= m, \\ m_5 &= 2m. \end{aligned}$$

The hypotheses are equally likely and the criterion is minimum Pr (ϵ).

1. Draw the decision regions on the R-axis.
2. Compute the error probability.

Problem 2.3.4. The observed random variable r has a Gaussian density on the three hypotheses,

$$p_{r|H_k}(R|H_k) = \frac{1}{\sqrt{2\pi}\,\sigma_k} \exp\left[-\frac{(R-m_k)^2}{2\sigma_k^2}\right], \qquad \begin{aligned} &-\infty < R < \infty \\ &k = 1, 2, 3, \end{aligned}$$

where the parameter values on the three hypotheses are,

$$\begin{aligned} &H_1: m_1 = 0, \quad \sigma_1 = \sigma_\alpha, \\ &H_2: m_2 = m, \quad \sigma_2 = \sigma_\alpha, \qquad (m > 0), \\ &H_3: m_3 = 0, \quad \sigma_3 = \sigma_\beta, \qquad (\sigma_\beta > \sigma_\alpha). \end{aligned}$$

The three hypotheses are equally likely and the criterion is minimum Pr (ϵ).

1. Find the optimum Bayes test.
2. Draw the decision regions on the R-axis for the special case,

$$\begin{aligned} \sigma_\beta^2 &= 2\sigma_\alpha^2, \\ \sigma_\alpha &= m. \end{aligned}$$

3. Compute the Pr (ϵ) for this special case.

Problem 2.3.5. The probability density of $\mathbf{r}$ on the three hypotheses is

$$p_{r_1,r_2|H_k}(R_1, R_2|H_k) = (2\pi\sigma_{1k}\sigma_{2k})^{-1} \exp\left[-\frac{1}{2}\left(\frac{R_1^2}{\sigma_{1k}^2} + \frac{R_2^2}{\sigma_{2k}^2}\right)\right], \qquad \begin{aligned} &-\infty < R_1, R_2 < \infty, \\ &k = 1, 2, 3, \end{aligned}$$

where

$$\begin{aligned} \sigma_{11}^2 &= \sigma_{21}^2 = \sigma_n^2, \\ \sigma_{12}^2 &= \sigma_s^2 + \sigma_n^2, \qquad \sigma_{22}^2 = \sigma_n^2, \\ \sigma_{13}^2 &= \sigma_n^2, \qquad \sigma_{23}^2 = \sigma_s^2 + \sigma_n^2. \end{aligned}$$

The cost matrix is

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & \alpha \\ 1 & \alpha & 0 \end{bmatrix},$$

where $0 \leq \alpha < 1$ and $\Pr(H_2) = \Pr(H_3) \triangleq p$. Define $l_1 = R_1^2$ and $l_2 = R_2^2$.

1. Find the optimum test and indicate the decision regions in the l_1, l_2-plane.
2. Write an expression for the error probabilities. (Do not evaluate the integrals.)
3. Verify that for $\alpha = 0$ this problem reduces to 2.2.17.

Problem 2.3.6. On H_k the observation is a value of a Poisson random variable

$$\Pr(r = n) = \frac{k_m^n}{n!} e^{-k_m}, \qquad m = 1, 2, \ldots, M,$$

where $k_m = mk$. The hypotheses are equally likely and the criterion is minimum Pr (ϵ).

1. Find the optimum test.
2. Find a simple expression for the boundaries of the decision regions and indicate how you would compute the Pr (ϵ).

Problem 2.3.7. Assume that the received vector on each of the three hypotheses is

$$\begin{aligned} H_0&: \mathbf{r} = \mathbf{m}_0 + \mathbf{n}, \\ H_1&: \mathbf{r} = \mathbf{m}_1 + \mathbf{n}, \\ H_2&: \mathbf{r} = \mathbf{m}_2 + \mathbf{n}, \end{aligned}$$

where

$$\mathbf{r} \triangleq \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix}, \qquad \mathbf{m}_i \triangleq \begin{bmatrix} m_{i1} \\ m_{i2} \\ m_{i3} \end{bmatrix}, \qquad \mathbf{n} \triangleq \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}.$$

The $\mathbf{m}_i$ are known vectors, and the components of $\mathbf{n}$ are statistically independent, zero-mean Gaussian random variables with variance σ^2.

1. Using the results in the text, express the Bayes test in terms of two sufficient statistics.

$$l_1 = \sum_{i=1}^{3} c_i r_i,$$

$$l_2 = \sum_{i=1}^{3} d_i r_i.$$

Find explicit expressions for c_i and d_i. Is the solution unique?

2. Sketch the decision regions in the l_1, l_2-plane for the particular cost assignment,

$$\begin{aligned} C_{00} &= C_{11} = C_{22} = 0, \\ C_{12} &= C_{21} = C_{01} = C_{10} = \tfrac{1}{2}C_{02} = \tfrac{1}{2}C_{20} > 0. \end{aligned}$$

P2.4 Estimation

Bayes Estimation

Problem 2.4.1. Let

$$r = ab + n,$$

where a, b, and n are independent zero-mean Gaussian variables with variances σ_a^2, σ_b^2, and σ_n^2.

1. What is $\hat{a}_{\text{map}}$?
2. Is this equivalent to simultaneously finding $\hat{a}_{\text{map}}$, $\hat{b}_{\text{map}}$?

3. Now consider the case in which

$$r = a + \sum_{i=1}^{k} b_i + n,$$

where the b_i are independent zero-mean Gaussian variables with variances $\sigma_{b_i}^2$.

(a) What is $\hat{a}_{\text{map}}$?
(b) Is this equivalent to simultaneously finding $\hat{a}_{\text{map}}$, $\hat{b}_{i,\text{map}}$?
(c) Explain intuitively why the answers to part 2 and part 3b are different.

Problem 2.4.2. The observed random variable is x. We want to estimate the parameter λ. The probability density of x as a function of λ is,

$$\begin{aligned} p_{x|\lambda}(X|\lambda) &= \lambda e^{-\lambda X}, & X \geq 0,\ \lambda > 0, \\ &= 0, & X < 0. \end{aligned}$$

The a priori density of λ depends on two parameters: n_*, l_*.

$$p_{\lambda|n_*,l_*}(\lambda|n_*, l_*) \triangleq \begin{cases} \dfrac{l_*^{n_*}}{\Gamma(n_*)} e^{-\lambda l_*} \lambda^{n_*-1}, & \lambda \geq 0, \\ 0, & \lambda < 0. \end{cases}$$

1. Find $E(\lambda)$ and Var (λ) *before* any observations are made.
2. Assume that one observation is made. Find $p_{\lambda|x}(\lambda|X)$. What interesting property does this density possess? Find $\hat{\lambda}_{\text{ms}}$ and $E[(\hat{\lambda}_{\text{ms}} - \lambda)^2]$.
3. Now assume that n independent observations are made. Denote these n observations by the vector $\mathbf{x}$. Verify that

$$p_{\lambda|\mathbf{x}}(\lambda|\mathbf{X}) \triangleq \begin{cases} \dfrac{(l')^{n'}}{\Gamma(n')} e^{-\lambda l'} \lambda^{n'-1}, & \lambda \geq 0, \\ 0, & \lambda < 0, \end{cases}$$

where

$$\begin{aligned} l' &= l + l_*, \\ n' &= n + n_*, \end{aligned}$$

and

$$l = \sum_{i=1}^{n} X_i.$$

Find $\hat{\lambda}_{\text{ms}}$ and $E[(\hat{\lambda}_{\text{ms}} - \lambda)^2]$.
4. Does $\hat{\lambda}_{\text{map}} = \hat{\lambda}_{\text{ms}}$?

Comment. Reproducing Densities. The reason that the preceding problem was simple was that the a priori and a posteriori densities had the same functional form. (Only the parameters changed.) In general,

$$p_{a|r}(A|R) = \frac{p_{r|a}(R|A)p_a(A)}{p_r(R)},$$

and we say that $p_a(A)$ is a *reproducing density* or a *conjugate prior* density [with respect to the transition density $p_{r|a}(R|A)$] if the a posteriori density is of the same form as $p_a(A)$. Because the choice of the a priori density is frequently somewhat arbitrary, it is convenient to choose a reproducing density in many cases. The next two problems illustrate other reproducing densities of interest.

Problem 2.4.3. Let

$$r = a + n,$$

where n is $N(0, \sigma_n)$. Then

$$p_{r|a}(R|A) = \frac{1}{\sqrt{2\pi}\,\sigma_n} \exp\left[-\frac{(R-A)^2}{2\sigma_n^2}\right].$$

1. Verify that a conjugate priori density for a is $N\left(m_0, \dfrac{\sigma_n}{k_0}\right)$ by showing that

$$p_{a|r}(A|R) = N(m_1, \sigma_1),$$

where

$$m_1 = \frac{m_0 k_0^2 + R}{(1 + k_0^2)}$$

and

$$\sigma_1^2 = \frac{\sigma_n^2}{1 + k_0^2}.$$

2. Extend this result to N independent observations by verifying that

$$p_{a|\mathbf{r}}(A|\mathbf{R}) = N(m_N, \sigma_N),$$

where

$$m_N = \frac{m_0 k_0^2 + Nl}{N + k_0^2},$$

$$\sigma_N^2 = \frac{\sigma_n^2}{N + k_0^2},$$

and

$$l \triangleq \frac{1}{N} \sum_{i=1}^{N} R_i.$$

Observe that the a priori parameter k_0 can be interpreted as an equivalent number of observations (fractional observations are allowed).

Problem 2.4.4. Consider the observation process

$$p_{r|a}(R|A) = \frac{A^{1/2}}{(2\pi)^{1/2}} \exp\left[-\frac{A}{2}(R - m)^2\right],$$

where m is known and A is the parameter of interest (it is the reciprocal of the standard deviation). We assume that N independent observations are available.

1. Verify that

$$p_a(A|k_1, k_2) = c(A^{k_1/2-1}) \exp\left(-\tfrac{1}{2}Ak_1k_2\right), \qquad \begin{aligned} A &\geq 0, \\ k_1, k_2 &> 0, \end{aligned}$$

(c is a normalizing factor) is a conjugate prior density by showing that

$$p_{a|\mathbf{r}}(A|\mathbf{R}) = p_a(A|k_1', k_2'),$$

where

$$\begin{aligned} k_2' &= \frac{1}{k_1'}(k_1k_2 + Nw), \\ k_1' &= k_1 + N, \\ w &= \frac{1}{N}\sum_{i=1}^{N}(R_i - m)^2. \end{aligned}$$

Note that k_1, k_2 are simply the parameters in the a priori density which are chosen based on our a priori knowledge.

2. Find $\hat{a}_{\text{ms}}$.

Problem 2.4.5. We make K observations: $R_1, \dots, R_K$, where

$$r_i = a + n_i.$$

The random variable a has a Gaussian density $N(0, \sigma_a)$. The n_i are independent Gaussian variables $N(0, \sigma_n)$.

1. Find the MMSE estimate $\hat{a}_{ms}$.
2. Find the MAP estimate $\hat{a}_{map}$.
3. Compute the mean-square error.
4. Consider an alternate procedure using the same r_i.

(a) Estimate a after each observation using a MMSE criterion.

This gives a sequence of estimates $\hat{a}_1(R_1)$, $\hat{a}_2(R_1, R_2) \dots \hat{a}_j(R_1, \dots R_j) \dots$ $\hat{a}_K(R_1, \dots, R_K)$. Denote the corresponding variances as $\sigma_1^2, \sigma_2^2, \dots, \sigma_K^2$.

(b) Express $\hat{a}_j$ as a function of $\hat{a}_{j-1}$, σ_{j-1}^2, and R_j.

(c) Show that

$$\frac{1}{\sigma_j^2} = \frac{1}{\sigma_a^2} + \frac{j}{\sigma_n^2}.$$

Problem 2.4.6. [36]. In this problem we outline the proof of Property 2 on p. 61. The assumptions are the following:

(a) The cost function is a symmetric, nondecreasing function. Thus

$$C(X) = C(-X)$$

$$C(X_1) \geq C(X_2) \quad \text{for} \quad X_1 \geq X_2 \geq 0, \tag{P.1}$$

which implies

$$\frac{dC(X)}{dX} \geq 0 \quad \text{for} \quad X \geq 0. \tag{P.2}$$

(b) The a posteriori probability density is symmetric about its conditional mean and is nonincreasing.

(c)

$$\lim_{X \to \infty} C(X) p_{x|\mathbf{r}}(X|\mathbf{R}) = 0. \tag{P.3}$$

We use the same notation as in Property 1 on p. 61. Verify the following steps:

1. The conditional risk using the estimate $\hat{a}$ is

$$\mathcal{R}(\hat{a}|\mathbf{R}) = \int_{-\infty}^{\infty} C(Z) p_{z|\mathbf{r}}(Z + \hat{a} - \hat{a}_{ms}|\mathbf{R})\, dZ. \tag{P.4}$$

2. The difference in conditional risks is

$$\Delta\mathcal{R} = \mathcal{R}(\hat{a}|\mathbf{R}) - \mathcal{R}(\hat{a}_{ms}|\mathbf{R}) = \int_0^{\infty} C(Z)[p_{z|\mathbf{r}}(Z + \hat{a} - \hat{a}_{ms}|\mathbf{R}) p_{z|\mathbf{r}}(Z - \hat{a} + \hat{a}_{ms}|\mathbf{R}) - 2p_{z|\mathbf{r}}(Z|\mathbf{R})]\, dZ. \tag{P.5}$$

3. For $\hat{a} > \hat{a}_{ms}$ the integral of the terms in the bracket with respect to Z from 0 to Z_0 is

$$\int_0^{\hat{a} - \hat{a}_{ms}} [p_{z|\mathbf{r}}(Z_0 + Y|\mathbf{R}) - p_{z|\mathbf{r}}(Z_0 - Y|\mathbf{R})]\, dY \triangleq g(Z_0) \tag{P.6}$$

4. Integrate (P.5) by parts to obtain

$$\Delta\mathcal{R} = C(Z) g(Z) \Big|_0^{\infty} - \int_0^{\infty} \frac{dC(Z)}{dZ} g(Z)\, dZ, \qquad \hat{a} > \hat{a}_{ms}. \tag{P.7}$$

5. Show that the assumptions imply that the first term is zero and the second term is nonnegative.

6. Repeat Steps 3 to 5 with appropriate modifications for $\hat{a} < \hat{a}_{ms}$.

7. Observe that these steps prove that $\hat{a}_{ms}$ minimizes the Bayes risk under the above assumptions. Under what conditions will the Bayes estimate be unique?

Nonrandom Parameter Estimation

Problem 2.4.7. We make n statistically independent observations: $r_1, r_2, \dots, r_n$, with mean m and variance σ^2. Define the sample variance as

$$V = \frac{1}{n} \sum_{j=1}^{n} \left(R_j - \sum_{i=1}^{n} \frac{R_i}{n}\right)^2.$$

Is it an unbiased estimator of the actual variance?

Problem 2.4.8. We want to estimate a in a binomial distribution by using n observations.

$$\Pr(r \text{ events}|a) = \binom{n}{r} a^r (1 - a)^{n-r}, \qquad r = 0, 1, 2, \dots, n.$$

1. Find the ML estimate of a and compute its variance.
2. Is it efficient?

Problem 2.4.9.

1. Does an efficient estimate of the standard deviation σ of a zero-mean Gaussian density exist?
2. Does an efficient estimate of the variance σ^2 of a zero-mean Gaussian density exist?

Problem 2.4.10 (continuation). The results of Problem 2.4.9 suggest the general question. Consider the problem of estimating some function of the parameter A, say, $f_1(A)$. The observed quantity is R and $p_{r|a}(R|A)$ is known. Assume that A is a nonrandom variable.

1. What are the conditions for an efficient estimate $\hat{f_1}(A)$ to exist?
2. What is the lower bound on the variance of the error of any unbiased estimate of $f_1(A)$?
3. Assume that an efficient estimate of $f_1(A)$ exists. When can an efficient estimate of some other function $f_2(A)$ exist?

Problem 2.4.11. The probability density of r, given A_1 and A_2 is:

$$p_{r|a_1,a_2}(R|A_1, A_2) = (2\pi A_2)^{-\frac{1}{2}} \exp\left[-\frac{(R - A_1)^2}{2A_2}\right];$$

that is, A_1 is the mean and A_2 is the variance.

1. Find the joint ML estimates of A_1 and A_2 by using n independent observations.
2. Are they biased?
3. Are they coupled?
4. Find the error covariance matrix.

Problem 2.4.12. We want to transmit two parameters, A_1 and A_2. In a simple attempt to achieve a secure communication system we construct two signals to be transmitted over separate channels.

$$s_1 = x_{11} A_1 + x_{12} A_2,$$
$$s_2 = x_{21} A_1 + x_{22} A_2,$$

where x_{ij}, $i, j = 1, 2$, are known. The received variables are

$$r_1 = s_1 + n_1,$$
$$r_2 = s_2 + n_2.$$

The additive noises are independent, identically distributed, zero-mean Gaussian random variables, $N(0, \sigma_n)$. The parameters A_1 and A_2 are nonrandom.

1. Are the ML estimates $\hat{a}_1$ and $\hat{a}_2$ unbiased?
2. Compute the variance of the ML estimates $\hat{a}_1$ and $\hat{a}_2$.
3. Are the ML estimates efficient? In other words, do they satisfy the Cramér-Rao bound with equality?

Problem 2.4.13. Let

$$y = \sum_{i=1}^{N} x_i,$$

where the x_i are independent, zero-mean Gaussian random variables with variance σ_x^2. We observe y. In parts 1 through 4 treat N as a continuous variable.

1. Find the maximum likelihood estimate of N.
2. Is $\hat{n}_{ml}$ unbiased?
3. What is the variance of $\hat{n}_{ml}$?
4. Is $\hat{n}_{ml}$ efficient?
5. Discuss qualitatively how you would modify part 1 to take into account that N is discrete.

Problem 2.4.14. We observe a value of the discrete random variable x.

$$\Pr(x = i|A) = \frac{A^i}{i!} e^{-A}, \qquad i = 0, 1, 2, \ldots,$$

where A is nonrandom.

1. What is the lower bound on the variance of any unbiased estimate, $\hat{a}(x)$?
2. Assuming n independent observations, find an $\hat{a}(\mathbf{x})$ that is efficient.

Problem 2.4.15. Consider the Cauchy distribution

$$p_{x|a}(X|A) = \{\pi[1 + (X - A)^2]\}^{-1}.$$

Assume that we make n independent observations in order to estimate A.

1. Use the Cramér-Rao inequality to show that the variance of any unbiased estimate of A has a variance greater than $2/n$.
2. Is the sample mean a consistent estimate?
3. We can show that the sample median is asymptotically normal, $N(A, \pi/\sqrt{4n})$. (See pp. 367–369 of Cramér [9].) What is the asymptotic efficiency of the sample median as an estimator?

Problem 2.4.16. Assume that

$$p_{r_1,r_2|\rho}(R_1, R_2|\rho) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{-\frac{(R_1^2 - 2\rho R_1 R_2 + R_2^2)}{2(1-\rho^2)}\right\}.$$

We want to estimate the correlation coefficient ρ by using n independent observations of (R_1, R_2).

1. Find the equation for the ML estimate $\hat{\rho}$.
2. Find a lower bound on the variance of any unbiased estimate of ρ.

Mathematical Properties

Problem 2.4.17. Consider the biased estimate $\hat{a}(\mathbf{R})$ of the *nonrandom* parameter A.

$$E(\hat{a}(\mathbf{R})) = A + B(A).$$

Show that

$$E[(\hat{a}(\mathbf{R}) - A)^2] \geq \frac{(1 + dB(A)/dA)^2}{E\left\{\left[\frac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A}\right]^2\right\}}.$$

This is the Cramér-Rao inequality for biased estimates. Note that it is a bound on the mean-square error.

Problem 2.4.18. Let $p_{\mathbf{r}|a}(\mathbf{R}|A)$ be the probability density of $\mathbf{r}$, given A. Let h be an arbitrary random variable that is independent of $\boldsymbol{r}$ defined so that $A + h$ ranges over all possible values of A. Assume that $p_{h_1}(H)$ and $p_{h_2}(H)$ are *two* arbitrary probability densities for h. Assuming that $\hat{a}(\mathbf{R})$ is unbiased, we have

$$\int [\hat{a}(\mathbf{R}) - (A + H)] p_{\mathbf{r}|a}(\mathbf{R}|A + H)\, d\mathbf{R} = 0.$$

Multiplying by $p_{h_i}(H)$ and integrating over H, we have

$$\int dH\, p_{h_i}(H) \int [\hat{a}(\mathbf{R}) - (A + H)] p_{\mathbf{r}|a}(\mathbf{R}|A + H)\, d\mathbf{R} = 0.$$

1. Show that

$$\operatorname{Var}[\hat{a}(R) - A] \geq \frac{[E_1(h) - E_2(h)]^2}{\int \left(\frac{p_{\mathbf{r}|a}(\mathbf{R}|A + H)\left\{\int [p_{h_1}(H) - p_{h_2}(H)]\, dH\right\}^2}{p_{\mathbf{r}|a}(\mathbf{R}|A)}\right) d\mathbf{R}}$$

for any $p_{h_1}(H)$ and $p_{h_2}(H)$. Observe that because this is true for all $p_{h_1}(H)$ and $p_{h_2}(H)$, we may write

$$\operatorname{Var}[\hat{a}(R) - A] \geq \sup_{p_{h_1}, p_{h_2}} \text{(right-hand side of above equation).}$$

Comment. Observe that this bound does not require any regularity conditions. Barankin [15] has shown that this is the greatest lower bound.

Problem 2.4.19 (continuation). We now derive two special cases.

1. First, let $p_{h_2}(H) = \delta(H)$. What is the resulting bound?
2. Second, let $p_{h_1}(H) = \delta(H - H_0)$, where $H_0 \neq 0$. Show that

$$\operatorname{Var}[\hat{a}(\mathbf{R}) - A] \geq \left(\inf_{H_0}\left\{\frac{1}{H_0^2}\left[\int \frac{p^2_{\mathbf{r}|a}(\mathbf{R}|A + H_0)}{p_{\mathbf{r}|a}(\mathbf{R}|A)}\, d\mathbf{R} - 1\right]\right\}\right)^{-1}.$$

The infimum being over all $H_0 \neq 0$ such that $p_{\mathbf{r}|a}(\mathbf{R}|A) = 0$ implies

$$p_{\mathbf{r}|a}(\mathbf{R}|A + H_0) = 0.$$

3. Show that the bound given in part 2 is always as good as the Cramér-Rao inequality when the latter applies.

Problem 2.4.20. Let

$$\mathbf{a} = \mathbf{L}\mathbf{b},$$

where $\mathbf{L}$ is a nonsingular matrix and $\mathbf{a}$ and $\mathbf{b}$ are vector random variables. Prove that

$$\hat{\mathbf{a}}_{map} = \mathbf{L}\hat{\mathbf{b}}_{map} \qquad \text{and} \qquad \hat{\mathbf{a}}_{ms} = \mathbf{L}\hat{\mathbf{b}}_{ms}.$$

Problem 2.4.21. An alternate way to derive the Cramér-Rao inequality is developed in this problem. First, construct the vector $\mathbf{z}$.

$$\mathbf{z} \triangleq \begin{bmatrix} \hat{a}(\mathbf{R}) - A \\ \hdashline \dfrac{\partial \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A} \end{bmatrix}.$$

1. Verify that for unbiased estimates $E(\mathbf{z}) = \mathbf{0}$.

2. Assuming that $E(\mathbf{z}) = \mathbf{0}$, the covariance matrix is

$$\mathbf{\Lambda_z} = E(\mathbf{z}\mathbf{z}^T).$$

Using the fact that $\mathbf{\Lambda_z}$ is nonnegative definite, derive the Cramér-Rao inequality. If the equality holds, what does this imply about $|\mathbf{\Lambda_z}|$?

Problem 2.4.22. Repeat Problem 2.4.21 for the case in which a is a random variable. Define

$$\mathbf{z} = \left[\begin{array}{c} \hat{a}(\mathbf{R}) - a \\ \hdashline \dfrac{\partial \ln p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A} \end{array}\right]$$

and proceed as before.

Problem 2.4.23. Bhattacharyya Bound. Whenever an efficient estimate does not exist, we can improve on the Cramér-Rao inequality. In this problem we develop a conceptually simple but algebraically tedious bound for unbiased estimates of nonrandom variables.

1. Define an $(N + 1)$-dimensional vector,

$$\mathbf{z} \triangleq \left[\begin{array}{c} \hat{a}(\mathbf{R}) - A \\ \hdashline \dfrac{1}{p_{\mathbf{r}|a}(\mathbf{R}|A)} \dfrac{\partial p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A} \\ \hdashline \dfrac{1}{p_{\mathbf{r}|a}(\mathbf{R}|A)} \dfrac{\partial^2 p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^2} \\ \hdashline \vdots \\ \hdashline \dfrac{1}{p_{\mathbf{r}|a}(\mathbf{R}|A)} \dfrac{\partial^N p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^N} \end{array}\right].$$

Verify that

$$\mathbf{\Lambda_z} \triangleq E(\mathbf{z}\mathbf{z}^T) = \left[\begin{array}{c:c:c} \sigma_\epsilon^2 & 1 & \mathbf{0} \\ \hdashline 1 & & \\ \hdashline & \multicolumn{2}{c}{\tilde{\mathbf{J}}} \\ \mathbf{0} & & \end{array}\right]$$

What are the elements in $\tilde{\mathbf{J}}$? Is $\mathbf{\Lambda_z}$ nonnegative definite? Assume that $\tilde{\mathbf{J}}$ is positive definite. When is $\mathbf{\Lambda_z}$ *not* positive definite?

2. Verify that the results in part 1 imply

$$\sigma_\epsilon^2 \geq \tilde{J}^{11}.$$

This is the Bhattacharyya bound. Under what conditions does the equality hold?

3. Verify that for $N = 1$ the Bhattacharyya bound reduces to Cramér-Rao inequality.

4. Does the Bhattacharyya bound always improve as N increases?

Comment. In part 2 the condition for equality is

$$\hat{a}(\mathbf{R}) - A = \sum_{i=1}^{N} c_i(A) \frac{1}{p_{\mathbf{r}|a}(\mathbf{R}|A)} \frac{\partial^i p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^i}.$$

This condition could be termed Nth-order efficiency but does not seem to occur in many problems of interest.

5. Frequently it is easier to work with

$$\frac{\partial^i \ln p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^i}.$$

Rewrite the elements $\tilde{J}_{ij}$ in terms of expectations of combinations of these quantities for $N = 1$ and 2.

Problem 2.4.24 (continuation). Let $N = 2$ in the preceding problem.

1. Verify that

$$\sigma_\epsilon^2 \geq \frac{1}{\tilde{J}_{11}} + \frac{\tilde{J}_{12}^2}{\tilde{J}_{11}(\tilde{J}_{11}\tilde{J}_{22} - \tilde{J}_{12}^2)}.$$

The second term represents the improvement in the bound.

2. Consider the case in which $\mathbf{r}$ consists of M independent observations with identical densities and finite conditional means and variances. Denote the elements of $\tilde{\mathbf{J}}$ due to M observations as $\tilde{J}_{ij}(M)$. Show that $\tilde{J}_{11}(M) = M\tilde{J}_{11}(1)$. Derive similar relations for $\tilde{J}_{12}(M)$ and $\tilde{J}_{22}(M)$. Show that

$$\sigma_\epsilon^2 \geq \frac{1}{M\tilde{J}_{11}(1)} + \frac{\tilde{J}_{12}^2(1)}{2M^2\tilde{J}_{11}^4(1)} + o\left(\frac{1}{M^2}\right).$$

Problem 2.4.25. [11] Generalize the result in Problem 2.4.23 to the case in which we are estimating a function of A, say $f(A)$. Assume that the estimate is unbiased. Define

$$\mathbf{z} = \left[\begin{array}{c} \hat{a}(\mathbf{R}) - f(A) \\ \hdashline k_1 \dfrac{1}{p_{\mathbf{r}|a}(\mathbf{R}|A)} \dfrac{\partial p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A} \\ \hdashline k_2 \dfrac{1}{p_{\mathbf{r}|a}(\mathbf{R}|A)} \dfrac{\partial^2 p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^2} \\ \hdashline \vdots \\ \hdashline k_N \dfrac{1}{p_{\mathbf{r}|a}(\mathbf{R}|A)} \dfrac{\partial^N p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^N} \end{array}\right].$$

Let

$$y = [\hat{a}(\mathbf{R}) - f(A)] - \sum_{i=1}^{N} k_i \frac{1}{p_{\mathbf{r}|a}(\mathbf{R}|A)} \cdot \frac{\partial^i p_{\mathbf{r}|a}(\mathbf{R}|A)}{\partial A^i}.$$

1. Find an expression for $\xi_y = E[y^2]$. Minimize ξ_y by choosing the k_i appropriately.
2. Using these values of k_i, find a bound on $\mathrm{Var}[\hat{a}(\mathbf{R}) - f(A)]$.
3. Verify that the result in Problem 2.4.23 is obtained by letting $f(A) = A$ in (2).

Problem 2.4.26.

1. Generalize the result in Problem 2.4.23 to establish a bound on the mean-square error in estimating a random variable.
2. Verify that the matrix of concern is

$$\mathbf{\Lambda_z} = \left[\begin{array}{c:c:c} E(a_\epsilon^2) & 1 & \mathbf{0} \\ \hdashline 1 & & \\ \hdashline & \multicolumn{2}{c}{\tilde{\mathbf{J}}_T} \\ \mathbf{0} & & \end{array}\right].$$

What are the elements in $\tilde{\mathbf{J}}_T$?

3. Find $\boldsymbol{\Lambda}_{\mathbf{z}}$ for the special case in which a is $N(0, \sigma_a)$.

Multiple Parameters

Problem 2.4.27. In (239) we defined the partial derivative matrix $\nabla_{\mathbf{x}}$.

$$\nabla_{\mathbf{x}} \triangleq \begin{bmatrix} \dfrac{\partial}{\partial x_1} \\ \dfrac{\partial}{\partial x_2} \\ \vdots \\ \dfrac{\partial}{\partial x_n} \end{bmatrix}.$$

Verify the following properties.

1. The matrix $\mathbf{A}$ is $n \times 1$ and the matrix $\mathbf{B}$ is $n \times 1$. Show that

$$\nabla_{\mathbf{x}}(\mathbf{A}^T\mathbf{B}) = (\nabla_{\mathbf{x}}\mathbf{A}^T)\mathbf{B} + (\nabla_{\mathbf{x}}\mathbf{B}^T)\mathbf{A}.$$

2. If the $n \times 1$ matrix $\mathbf{B}$ is not a function of $\mathbf{x}$, show that

$$\nabla_{\mathbf{x}}(\mathbf{B}^T\mathbf{x}) = \mathbf{B}.$$

3. Let $\mathbf{C}$ be an $n \times m$ constant matrix,

$$\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{C}) = \mathbf{C}.$$

4. $\nabla_{\mathbf{x}}(\mathbf{x}^T) = \mathbf{I}$.

Problem 2.4.28. A problem that occurs frequently is the differentiation of a quadratic form.

$$Q = \mathbf{A}^T(\mathbf{x})\, \boldsymbol{\Lambda}\mathbf{A}(\mathbf{x}),$$

where $\mathbf{A}(\mathbf{x})$ is a $m \times 1$ matrix whose elements are a function of $\mathbf{x}$ and $\boldsymbol{\Lambda}$ is a symmetric nonnegative definite $m \times m$ matrix. Recall that this implies that we can write

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}.$$

1. Prove

$$\nabla_{\mathbf{x}}Q = 2(\nabla_{\mathbf{x}}\mathbf{A}^T(\mathbf{x}))\, \boldsymbol{\Lambda}\mathbf{A}(\mathbf{x})$$

2. For the special case

$$\mathbf{A}(\mathbf{x}) = \mathbf{B}\mathbf{x},$$

prove

$$\nabla_{\mathbf{x}}Q = 2\mathbf{B}^T\boldsymbol{\Lambda}\mathbf{B}\mathbf{x}.$$

3. For the special case

$$Q = \mathbf{x}^T\boldsymbol{\Lambda}\mathbf{x},$$

prove

$$\nabla_{\mathbf{x}}Q = 2\boldsymbol{\Lambda}\mathbf{x}.$$

Problem 2.4.29. Go through the details of the proof on p. 83 for arbitrary K.

Problem 2.4.30. As discussed in (284), we frequently estimate,

$$\mathbf{d} \triangleq \mathbf{g}_d(\mathbf{A}).$$

Assume the estimates are unbiased. Derive (286).

Problem 2.4.31. The cost function is a scalar-valued function of the vector $\mathbf{a}_\epsilon$, $C(\mathbf{a}_\epsilon)$. Assume that it is symmetric and convex,

1. $C(\mathbf{a}_\epsilon) = C(-\mathbf{a}_\epsilon)$,
2. $C(b\mathbf{x}_1 + (1 - b)\mathbf{x}_2) \leq bC(\mathbf{x}_1) + (1 - b)\, C(\mathbf{x}_2), \qquad 0 \leq b \leq 1.$

Assume that the a posteriori density is symmetric about its conditional mean. Prove that the conditional mean of **a** minimizes the Bayes risk.

Problem 2.4.32. Assume that we want to estimate K nonrandom parameters $A_1, A_2, \ldots, A_K$,denoted by **A**. The probability density $p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})$ is known. Consider the biased estimates $\hat{\mathbf{a}}(\mathbf{R})$ in which

$$\mathrm{B}(a_i) \triangleq \int [\hat{a}_i(\mathbf{R}) - A_i] p_{\mathbf{r}|\mathbf{a}}(\mathbf{R}|\mathbf{A})\, d\mathbf{R}.$$

1. Derive a bound on the mean-square error in estimating A_i.
2. The error correlation matrix is

$$\mathbf{R}_\epsilon \triangleq E[(\hat{\mathbf{a}}(\mathbf{R}) - \mathbf{A})(\hat{\mathbf{a}}^T(\mathbf{R}) - \mathbf{A}^T)]$$

Find a matrix $\mathbf{J}_B$ such that, $\mathbf{J}_B - \mathbf{R}_\epsilon^{-1}$ is nonnegative definite.

Miscellaneous

Problem 2.4.33. Another method of estimating nonrandom parameters is called the method of moments (Pearson [37]). If there are k parameters to estimate, the first k sample moments are equated to the actual moments (which are functions of the parameters of interest). Solving these k equations gives the desired estimates. To illustrate this procedure consider the following example. Let

$$\begin{aligned} p_{x|\lambda}(X|\lambda) &= \frac{1}{\Gamma(\lambda)} X^{\lambda-1}e^{-X}, \qquad && X \geq 0, \\ &= 0, && X < 0, \end{aligned}$$

where λ is a positive parameter. We have n independent observations of x.

1. Find a lower bound on the variance of any unbiased estimate.
2. Denote the method of moments estimate as $\hat{\lambda}_{mm}$. Show

$$\hat{\lambda}_{mm} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

and compute $E(\hat{\lambda}_{mm})$ and Var $(\hat{\lambda}_{mm})$.

Comment. In [9] the efficiency of $\hat{\lambda}_{mm}$ is computed. It is less than 1 and tends to zero as $n \to \infty$.

Problem 2.4.34. Assume that we have n independent observations from a Gaussian density $N(m, \sigma)$. Verify that the method of moments estimates of m and σ are identical to the maximum-likelihood estimates.

P2.5 Composite Hypotheses

Problem 2.5.1. Consider the following composite hypothesis testing problem,

$$H_0\!: p_r(R) = \frac{1}{\sqrt{2\pi}\, \sigma_0} \exp\left(-\frac{R^2}{2\sigma_0^2}\right),$$

where σ_0 is known,

$$H_1\!: p_r(R) = \frac{1}{\sqrt{2\pi}\, \sigma_1} \exp\left(-\frac{R^2}{2\sigma_1^2}\right),$$

where $\sigma_1 > \sigma_0$. Assume that we require $P_F = 10^{-2}$.

1. Construct an upper bound on the power function by assuming a perfect measurement scheme coupled with a likelihood ratio test.

2. Does a uniformly most powerful test exist?

3. If the answer to part 2 is negative, construct the power function of a generalized likelihood ratio test.

Problem 2.5.2. Consider the following composite hypothesis testing problem. *Two* statistically independent observations are received. Denote the observations as R_1 and R_2. Their probability densities on the two hypotheses are

$$H_0 : p_{r_i}(R_i) = \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left(-\frac{R_i^2}{2\sigma_0^2}\right), \qquad i = 1, 2,$$

where σ_0 is known,

$$H_1 : p_{r_i}(R_i) = \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left(-\frac{R_i^2}{2\sigma_1^2}\right), \qquad i = 1, 2,$$

where $\sigma_1 > \sigma_0$. Assume that we require a $P_F = \alpha$.

1. Construct an upper bound on the power function by assuming a perfect measurement scheme coupled with a likelihood ratio test.

2. Does a uniformly most powerful test exist?

3. If the answer to part 2 is negative, construct the power function of a generalized likelihood ratio test.

Problem 2.5.3. The observation consists of a set of values of the random variables, $r_1, r_2, \ldots, r_M$.

$$\begin{aligned} r_i &= s_i + n_i, \qquad & i &= 1, 2, \ldots, M, \qquad & &H_1, \\ r_i &= n_i, & i &= 1, 2, \ldots, M, & &H_0. \end{aligned}$$

The s_i and n_i are independent, identically distributed random variables with densities $N(0, \sigma_s)$ and $N(0, \sigma_n)$, respectively, where σ_n is known and σ_s is unknown.

1. Does a UMP test exist?

2. If the answer to part 1 is negative, find a generalized LRT.

Problem 2.5.4. The observation consists of a set of values of the random variables $r_1, r_2, \ldots, r_M$, which we denote by the vector $\mathbf{r}$. Under H_0 the r_i are statistically independent, with densities

$$p_{r_i}(R_i) = \frac{1}{\sqrt{2\pi\lambda_i^0}} \exp\left(-\frac{R_i^2}{2\lambda_i^0}\right)$$

in which the λ_i^0 are known. Under H_1 the r_i are statistically independent, with densities

$$p_{r_i}(R_i) = \frac{1}{\sqrt{2\pi\lambda_i^1}} \exp\left(-\frac{R_i^2}{2\lambda_i^1}\right)$$

in which $\lambda_i^1 > \lambda_i^0$ for all i. Repeat Problem 2.5.3.

Problem 2.5.5. Consider the following hypothesis testing problem. *Two* statistically independent observations are received. Denote the observations R_1 and R_2. The probability densities on the two hypotheses are

$$H_0 : p_{r_i}(R_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R_i^2}{2\sigma^2}\right), \qquad i = 1, 2,$$

$$H_1 : p_{r_i}(R_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(R_i - m)^2}{2\sigma^2}\right] \qquad i = 1, 2,$$

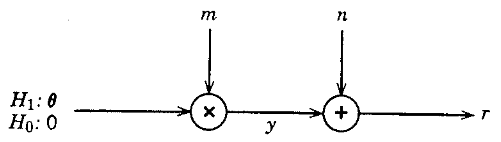
where m can be *any* nonzero number. Assume that we require $P_F = \alpha$.

1. Construct an upper bound on the power function by assuming a perfect measurement scheme coupled with a likelihood ratio test.

2. Does a uniformly most powerful test exist?

3. If the answer to part 2 is negative, construct the power function of a generalized likelihood ratio test.

Problem 2.5.6. Consider the following hypothesis-testing problem.



Under H_1 a nonrandom variable $\theta\,(-\infty < \theta < \infty)$ is transmitted. It is multiplied by the random variable m. A noise n is added to the result to give r. Under H_0 nothing is transmitted, and the output is just n. Thus

$$\begin{aligned} H_1 &: r = m\theta + n, \\ H_0 &: r = n. \end{aligned}$$

The random variables m and n are independent.

$$p_n(N) = \frac{1}{\sqrt{2\pi}\,\sigma_n} \exp\left(-\frac{N^2}{2\sigma_n^2}\right),$$

$$p_m(M) = \tfrac{1}{2}\,\delta(M - 1) + \tfrac{1}{2}\,\delta(M + 1).$$

1. Does a uniformly most powerful test exist? If it does, describe the test and give an expression for its power function? If it does not, indicate why.

2. Do one of the following:

(a) If a UMP test exists for this example, derive a necessary and sufficient condition on $p_m(M)$ for a UMP test to exist. (The rest of the model is unchanged.)

(b) If a UMP test does not exist, derive a generalized likelihood ratio test and an expression for its power function.

Problem 2.5.7 (CFAR receivers.) We have N independent observations of the variable x. The probability density on H_k is

$$p_{x_i|H_k}(X|H_k) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{\frac{-(X_i - m_k)^2}{2\sigma^2}\right\} \qquad -\infty < X_i < \infty, \qquad \begin{aligned} i &= 1, 2, \ldots N, \\ H_k &: k = 0,1, \\ m_0 &= 0. \end{aligned}$$

The variance σ^2 is *unknown*. Define

$$l_1 = \sum_{i=1}^{N} x_i$$

$$l_2 = \sum_{i=1}^{N} x_i^2$$

(a) Consider the test

$$l_1^2 \underset{H_0}{\overset{H_1}{\gtrless}} \alpha l_2$$

Verify that the P_F of this test does *not* depend on σ^2. (*Hint*. Use formula in Problem 2.4.6.)

(b) Find α as a function of P_F.

(c) Is this a UMP test?

(d) Consider the particular case in which $N = 2$ and $m_1 = m$. Find P_D as a function of P_F and m/σ. Compare your result with Figure 2.9*b* and see how much the lack of knowledge about the variance σ^2 has decreased the system performance.

Comment. Receivers of this type are called CFAR (constant false alarm rate) receivers in the radar/sonar literature.

Problem 2.5.8 (continuation). An alternate approach to the preceding problem would be a generalized LRT.

1. Find the generalized LRT and write an expression for its performance for the case in which $N = 2$ and $m_1 = m$.
2. How would you decide which test to use?

Problem 2.5.9. Under H_0, x is a Poisson variable with a known intensity k_0.

$$\Pr(x = n) = \frac{k_0^n}{n!} e^{-k_0}, \qquad n = 0, 1, 2, \ldots.$$

Under H_1, x is a Poisson variable with an unknown intensity k_1, where $k_1 > k_0$.

1. Does a UMP test exist?
2. If a UMP test does not exist, assume that M independent observations of x are available and construct a generalized LRT.

Problem 2.5.10. How are the results to Problem 2.5.2 changed if we know that $\sigma_0 < \sigma_c$ and $\sigma_1 > \sigma_c$ where σ_c is known. Neither σ_0 or σ_1, however, is known. If a UMP test does not exist, what test procedure (other than a generalized LRT) would be logical?

P2.6 General Gaussian Problem

Detection

Problem 2.6.1. The M-hypothesis, general Gaussian problem is

$$p_{\mathbf{r}|H_i}(\mathbf{R}|H_i) = [(2\pi)^{N/2}|\mathbf{K}_i|^{1/2}]^{-1} \exp\left[-\tfrac{1}{2}(\mathbf{R}^T - \mathbf{m}_i^T)\,\mathbf{Q}_i(\mathbf{R} - \mathbf{m}_i)\right], \qquad i = 1, 2, \ldots, M.$$

1. Use the results of Problem 2.3.2 to find the Bayes test for this problem.
2. For the particular case in which the cost of a correct decision is zero and the cost of any wrong decision is equal show that the test reduces to the following:

Compute

$$l_i(\mathbf{R}) = \ln P_i - \tfrac{1}{2}\ln|\mathbf{K}_i| - \tfrac{1}{2}(\mathbf{R}^T - \mathbf{m}_i^T)\,\mathbf{Q}_i(\mathbf{R} - \mathbf{m}_i)$$

and choose the largest.

Problem 2.6.2 (continuation). Consider the special case in which all $\mathbf{K}_i = \sigma_n^2\mathbf{I}$ and the hypotheses are equally likely. Use the costs in Part 2 of Problem 2.6.1.

1. What determines the dimension of the decision space? Draw some typical decision spaces to illustrate the various alternatives.
2. Interpret the processor as a minimum-distance decision rule.

Problem 2.6.3. Consider the special case in which $\mathbf{m}_i = 0$, $i = 1, 2, \ldots, M$, and the hypotheses are equally likely. Use the costs in Part 2 of Problem 2.6.1.

1. Show that the test reduces to the following:

Compute

$$l_i(\mathbf{R}) = \mathbf{R}^T\mathbf{Q}_i\mathbf{R} + \ln|\mathbf{K}_i|$$

and choose the *smallest*.

2. Write an expression for the $\Pr(\epsilon)$ in terms of $p_{l|H_k}(\mathbf{L}|H_k)$, where

$$l \triangleq \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_M \end{bmatrix}.$$

Problem 2.6.4. Let

$$q_{\mathbf{B}} \triangleq \mathbf{x}^T\mathbf{B}\mathbf{x},$$

where $\mathbf{x}$ is a Gaussian vector $N(\mathbf{0}, \mathbf{I})$ and $\mathbf{B}$ is a symmetric matrix.

1. Verify that the characteristic function of $q_{\mathbf{B}}$ is

$$M_{q_{\mathbf{B}}}(jv) \triangleq E(e^{jvq_{\mathbf{B}}}) = \prod_{i=1}^{N} (1 - 2jv\lambda_{\mathbf{B}i})^{-1/2},$$

where $\lambda_{\mathbf{B}i}$ are the eigenvalues of $\mathbf{B}$.

2. What is $p_{q_{\mathbf{B}}}(Q)$ when the eigenvalues are equal?
3. What is the form of $p_{q_{\mathbf{B}}}(Q)$ when N is even and the eigenvalues are pair-wise equal but otherwise distinct; that is,

$$\begin{aligned} \lambda_{2i-1} &= \lambda_{2i}, \qquad i = 1, 2, \ldots, \frac{N}{2}, \\ \lambda_{2i} &\neq \lambda_{2j}, \qquad \text{all } i \neq j. \end{aligned}$$

Problem 2.6.5.

1. Modify the result of the preceding problem to include the case in which $\mathbf{x}$ is a Gaussian vector $N(\mathbf{0}, \mathbf{\Lambda_x})$, where $\mathbf{\Lambda_x}$ is positive definite.
2. What is $M_{q_{\mathbf{\Lambda_x}^{-1}}}(jv)$? Does the result have any interesting features?

Problem 2.6.6. Consider the M-ary hypothesis-testing problem. *Each* observation is a three-dimensional vector.

$$\begin{aligned} H_0&: \mathbf{r} = \mathbf{m}_0 + \mathbf{n}, \\ H_1&: \mathbf{r} = \mathbf{m}_1 + \mathbf{n}, \\ H_2&: \mathbf{r} = \mathbf{m}_2 + \mathbf{n}, \\ H_3&: \mathbf{r} = \mathbf{m}_3 + \mathbf{n}, \\ \mathbf{m}_1 &= +A, 0, B, \\ \mathbf{m}_2 &= 0, +A, B, \\ \mathbf{m}_3 &= -A, 0, B, \\ \mathbf{m}_4 &= 0, -A, B. \end{aligned}$$

The components of the noise vector are independent, identically distributed Gaussian variables, $N(0, \sigma)$. We have K independent observations. Assume a minimum $\Pr(\epsilon)$ criterion and equally-likely hypotheses. Sketch the decision region and compute the $\Pr(\epsilon)$.

Problem 2.6.7. Consider the following detection problem. Under either hypothesis the observation is a *two*-dimensional vector $\mathbf{r}$.

Under H_1:

$$\mathbf{r} \triangleq \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} = \mathbf{x} + \mathbf{n}.$$

Under H_0:

$$\mathbf{r} \triangleq \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} = \mathbf{y} + \mathbf{n}.$$

The signal vectors **x** and **y** are known. The length of the signal vector is constrained to equal $\sqrt{E}$ under both hypotheses; that is,

$$x_1^2 + x_2^2 = E,$$
$$y_1^2 + y_2^2 = E.$$

The noises are *correlated* Gaussian variables.

$$p_{n_1 n_2}(N_1, N_2) = \frac{1}{2\pi\sigma^2(1-\rho^2)^{1/2}} \exp\left(-\frac{N_1^2 - 2\rho N_1 N_2 + N_2^2}{2\sigma^2(1-\rho^2)}\right).$$

1. Find a sufficient statistic for a likelihood ratio test. Call this statistic $l(\mathbf{R})$. We have already shown that the quantity

$$d^2 = \frac{[E(l|H_1) - E(l|H_0)]^2}{\text{Var}\,(l|H_0)}$$

characterizes the performance of the test in a monotone fashion.

2. Choose **x** and **y** to maximize d^2. Does the answer depend on ρ?
3. Call the d^2 obtained by using the best **x** and **y**, d_0^2. Calculate d_0^2 for $\rho = -1, 0,$ and draw a rough sketch of d_0^2 as ρ varies from -1 through 0 to 1.
4. Explain why the performance curve in part 3 is intuitively correct.

Estimation

Problem 2.6.8. The observation is an N-dimensional vector

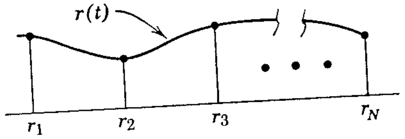
$$\mathbf{r} = \mathbf{a} + \mathbf{n},$$

where **a** is $N(\mathbf{0}, \mathbf{K_a})$, **n** is $N(\mathbf{0}, \mathbf{K_n})$, and **a** and **n** are statistically independent.

1. Find $\hat{\mathbf{a}}_{\text{map}}$. *Hint.* Use the properties of $\nabla_{\mathbf{a}}$ developed in Problems 2.4.27 and 2.4.28.
2. Verify that $\hat{\mathbf{a}}_{\text{map}}$ is efficient.
3. Compute the error correlation matrix

$$\mathbf{\Lambda}_\epsilon \triangleq E[(\hat{\mathbf{a}}_{\text{ms}} - \mathbf{a})(\hat{\mathbf{a}}_{\text{ms}} - \mathbf{a})^T].$$

Comment. Frequently this type of observation vector is obtained by sampling a random process $r(t)$ as shown below,



We denote the N samples by the vector **r**. Using **r**, we estimate the samples of $a(t)$ which are denoted by a_i. An error of interest is the sum of the squares of errors in estimating the a_i.

$$a_{\epsilon_i} = \hat{a}_i - a,$$

then

$$\xi_I \triangleq E\left[\sum_{i=1}^{N} (\hat{a}_i - a)^2\right] = E\left(\sum_{i=1}^{N} a_{\epsilon_i}{}^2\right) = E(\mathbf{a}_\epsilon{}^T\mathbf{a}_\epsilon) = \text{Tr}\,(\mathbf{\Lambda}_\epsilon).$$

Problem 2.6.9 (continuation). Consider the special case

$$\mathbf{K}_n = \sigma_n^2\mathbf{I}.$$

1. Verify that

$$\hat{\mathbf{a}}_{\text{ms}} = (\sigma_n^2\mathbf{I} + \mathbf{K}_a)^{-1}\mathbf{K}_a\mathbf{R}.$$

2. Now recall the detection problem described in Case 1 on p. 107. Verify that

$$l(\mathbf{R}) = \frac{1}{\sigma_n^2}\mathbf{R}^T\hat{\mathbf{a}}_{\text{ms}}.$$

Draw a block diagram of the processor. Observe that this is identical to the "unequal mean-equal covariance" case, except the mean **m** has been replaced by the mean-square estimate of the mean, $\hat{\mathbf{a}}_{ms}$.

3. What is the mean-square estimation error ξ_I?

Problem 2.6.10. Consider an alternate approach to Problem 2.6.8.

$$\mathbf{r} = \mathbf{a} + \mathbf{n},$$

where **a** is $N(\mathbf{0}, \mathbf{K_a})$ and **n** is $N(\mathbf{0}, \sigma_n^2\mathbf{I})$. Pass **r** through the matrix operation **W**, which is defined in (369). The eigenvectors are those of $\mathbf{K}_a$.

$$\mathbf{r}' \triangleq \mathbf{W}\mathbf{r} = \mathbf{x} + \mathbf{n}'$$

1. Verify that $\mathbf{W}\mathbf{W}^T = \mathbf{I}$.
2. What are the statistics of **x** and **n**′?
3. Find $\hat{\mathbf{x}}$. Verify that

$$\hat{x}_i = \frac{\lambda_i}{\lambda_i + \sigma_n^2} R_i',$$

where λ_i are the eigenvalues of $\mathbf{K_a}$.

4. Express $\hat{\mathbf{a}}$ in terms of a linear transformation of $\hat{\mathbf{x}}$. Draw a block diagram of the over-all estimator.
5. Prove

$$\xi_I \triangleq E[\mathbf{a}_\epsilon{}^T\mathbf{a}_\epsilon] = \sigma_n^2 \sum_{i=1}^{N} \frac{\lambda_i}{\lambda_i + \sigma_n^2}.$$

Problem 2.6.11 (Nonlinear Estimation). In the general Gaussian nonlinear estimation problem

$$\mathbf{r} = \mathbf{s}(\mathbf{A}) + \mathbf{n},$$

where $\mathbf{s}(\mathbf{A})$ is a nonlinear function of **A**. The noise **n** is Gaussian $N(\mathbf{0}, \mathbf{K_n})$ and independent of **a**.

1. Verify that

$$p_{\mathbf{r}|\mathbf{s}(\mathbf{A})}(\mathbf{R}|\mathbf{s}(\mathbf{A})) = [(2\pi)^{N/2}|\mathbf{K_n}|^{1/2}]^{-1} \exp\left[-\tfrac{1}{2}(\mathbf{R}^T - \mathbf{s}^T(\mathbf{A}))\mathbf{Q_n}\,(\mathbf{R} - \mathbf{s}(\mathbf{A}))\right].$$

2. Assume that **a** is a Gaussian vector $N(\mathbf{0}, \mathbf{K_a})$. Find an expression for $\ln p_{\mathbf{r},\mathbf{a}}(\mathbf{R}, \mathbf{A})$.
3. Using the properties of the derivative matrix $\nabla_{\mathbf{a}}$ derived in Problems 2.4.27 and 2.4.28, find the MAP equation.

Problem 2.6.12 (Optimum Discrete Linear Filter). Assume that we have a sequence of scalar observations $r_1, r_2, r_3, \ldots, r_K$, where $r_i = a_i + n_i$ and

$$E(a_i) = E(n_i) = 0,$$
$$E(\mathbf{r}\mathbf{r}^T) = \mathbf{\Lambda_r}, \qquad (N \times N),$$
$$E(\mathbf{r}a_i) = \mathbf{\Lambda}_{\mathbf{r}a_i}, \qquad (N \times 1).$$

We want to estimate a_K by using a realizable discrete linear filter. Thus

$$\hat{a}_K = \sum_{i=1}^{K} h_i R_i = \mathbf{h}^T\mathbf{R}.$$

Define the mean-square point estimation error as

$$\xi_P \triangleq E\{[\hat{a}_K(\mathbf{R}) - a_K]^2\}.$$

1. Use $\nabla_{\mathbf{h}}$ to find the discrete linear filter that minimizes ξ_P.
2. Find ξ_P for the optimum filter.
3. Consider the special case in which **a** and **n** are statistically independent. Find **h** and ξ_P.
4. How is $\hat{a}_K(\mathbf{R})$ for part 3 related to $\hat{\mathbf{a}}_{\text{map}}$ in Problem 2.6.8.

Note. No assumption about Gaussianness has been used.

Sequential Estimation

Problem 2.6.13. Frequently the observations are obtained in a time-sequence, $r_1, r_2, r_3, \ldots, r_N$. We want to estimate the k-dimensional parameter **a** in a sequential manner.

The ith observation is

$$r_i = \mathbf{Ca} + w_i, \qquad i = 1, 2, \ldots, N,$$

where **C** is a known $1 \times k$ matrix. The noises w_i are independent, identically distributed Gaussian variables $N(0, \sigma_n)$. The a priori knowledge is that **a** is Gaussian, $N(\mathbf{m}_0, \mathbf{\Lambda_a})$.

1. Find $p_{\mathbf{a}|r_1}(\mathbf{A}|R_1)$.
2. Find the minimum mean-square estimate $\hat{\mathbf{a}}_1$ and the error correlation matrix $\mathbf{\Lambda}_{\epsilon_1}$. Put your answer in the form

$$p_{\mathbf{a}|r_1}(\mathbf{A}|R_1) = c \exp\left[-\tfrac{1}{2}(\mathbf{A} - \hat{\mathbf{a}}_1)^T \mathbf{\Lambda}_{\epsilon_1}^{-1}(\mathbf{A} - \hat{\mathbf{a}}_1)\right],$$

where

$$\mathbf{\Lambda}_{\epsilon_1}^{-1} = \mathbf{\Lambda_a}^{-1} + \mathbf{C}^T\sigma_n^{-2}\mathbf{C}$$

and

$$\hat{\mathbf{a}}_1 = \mathbf{m}_0 + \frac{1}{\sigma_n^2}\mathbf{\Lambda}_{\epsilon_1}\mathbf{C}^T(R_1 - \mathbf{Cm}_0).$$

3. Draw a block diagram of the optimum processor.
4. Now proceed to the second observation R_2. What is the a priori density for this observation? Write the equations for $p_{\mathbf{a}|r_1,r_2}(\mathbf{A}|r_1, r_2)$, $\mathbf{\Lambda}_{\epsilon_2}^{-1}$, and $\hat{\mathbf{a}}_2$ in the same format as above.
5. Draw a block diagram of the sequential estimator and indicate exactly what must be stored at the end of each estimate.

Problem 2.6.14. Problem 2.6.13 can be generalized by allowing each observation to be an m-dimensional vector. The ith observation is

$$\mathbf{r}_i = \mathbf{Ca} + \mathbf{w}_i,$$

where **C** is a known $m \times k$ matrix. The noise vectors $\mathbf{w}_i$ are independent, identically distributed Gaussian vectors, $N(\mathbf{0}, \mathbf{\Lambda_w})$, where $\mathbf{\Lambda_w}$ is positive-definite.

Repeat Problem 2.6.13 for this model. Verify that

$$\hat{\mathbf{a}}_i = \hat{\mathbf{a}}_{i-1} + \mathbf{\Lambda}_{\epsilon_i}\mathbf{C}^T\mathbf{\Lambda_w}^{-1}(\mathbf{R}_i - \mathbf{C}\hat{\mathbf{a}}_{i-1})$$

and

$$\mathbf{\Lambda}_{\epsilon_i}^{-1} = \mathbf{\Lambda}_{\epsilon_{i-1}}^{-1} + \mathbf{C}^T\mathbf{\Lambda_w}^{-1}\mathbf{C}.$$

Draw a block diagram of the optimum processor.

***Problem 2.6.15.* Discrete Kalman Filter.** Now consider the case in which the parameter **a** changes according to the equation

$$\mathbf{a}_{k+1} = \mathbf{\Phi a}_k + \mathbf{\Gamma u}_k, \qquad k = 1, 2, 3, \ldots,$$

where $\mathbf{a}_1$ is $N(\mathbf{m}_0, \mathbf{P}_0)$, $\mathbf{\Phi}$ is an $n \times n$ matrix (known), $\mathbf{\Gamma}$ is an $n \times p$ matrix (known), $\mathbf{u}_k$ is $N(\mathbf{0}, \mathbf{Q})$, and $\mathbf{u}_k$ is independent of $\mathbf{u}_j$ for $j \neq k$. The observation process is

$$\mathbf{r}_k = \mathbf{Ca}_k + \mathbf{w}_k, \qquad k = 1, 2, 3, \ldots,$$

where **C** is an $m \times n$ matrix, $\mathbf{w}_k$ is $N(\mathbf{0}, \mathbf{\Lambda_w})$ and the $\mathbf{w}_k$ are independent of each other and $\mathbf{u}_j$.

Part I. We first estimate $\mathbf{a}_1$, using a mean-square error criterion.

1. Write $p_{\mathbf{a}_1|\mathbf{r}_1}(\mathbf{A}_1|\mathbf{R}_1)$.
2. Use the $\nabla_{\mathbf{a}_1}$ operator to obtain $\hat{\mathbf{a}}_1$.
3. Verify that $\hat{\mathbf{a}}_1$ is efficient.
4. Use $\nabla_{\mathbf{a}_1}\{[\nabla_{\mathbf{a}_1}(\ln p_{\mathbf{a}_1|\mathbf{r}_1}(\mathbf{A}_1|\mathbf{R}_1))]^T\}$ to find the error covariance matrix $\mathbf{P}_1$, where

$$\mathbf{P}_i \triangleq E[(\hat{\mathbf{a}}_i - \mathbf{a}_i)(\hat{\mathbf{a}}_i - \mathbf{a}_i)^T], \qquad i = 1, 2, \ldots.$$

Check.

$$\hat{\mathbf{a}}_1 = \mathbf{m}_0 + \mathbf{P}_1\mathbf{C}^T\mathbf{\Lambda_w}^{-1}[\mathbf{R} - \mathbf{Cm}_0]$$

and

$$\mathbf{P}_1^{-1} = \mathbf{P}_0^{-1} + \mathbf{C}^T\mathbf{\Lambda_w}^{-1}\mathbf{C}.$$

Part II. Now we estimate $\mathbf{a}_2$.

1. Verify that

$$p_{\mathbf{a}_2|\mathbf{r}_1,\mathbf{r}_2}(\mathbf{A}_2|\mathbf{R}_1, \mathbf{R}_2) = \frac{p_{\mathbf{r}_2|\mathbf{a}_2}(\mathbf{R}_2|\mathbf{A}_2)\, p_{\mathbf{a}_2|\mathbf{r}_1}(\mathbf{A}_2|\mathbf{R}_1)}{p_{\mathbf{r}_2|\mathbf{r}_1}(\mathbf{R}_2|\mathbf{R}_1)}.$$

2. Verify that $p_{\mathbf{a}_2|\mathbf{r}_1}(\mathbf{A}_2|\mathbf{R}_1)$ is $N(\mathbf{\Phi}\hat{\mathbf{a}}_1, \mathbf{M}_2)$, where

$$\mathbf{M}_2 \triangleq \mathbf{\Phi P}_1\mathbf{\Phi}^T + \mathbf{\Gamma Q \Gamma}^T.$$

3. Find $\hat{\mathbf{a}}_2$ and $\mathbf{P}_2$.

Check.

$$\hat{\mathbf{a}}_2 = \mathbf{\Phi}\hat{\mathbf{a}}_1 + \mathbf{P}_2\mathbf{C}^T\mathbf{\Lambda_w}^{-1}(\mathbf{R}_2 - \mathbf{C\Phi}\hat{\mathbf{a}}_1),$$
$$\mathbf{P}_2^{-1} = \mathbf{M}_2^{-1} + \mathbf{C}^T\mathbf{\Lambda_w}^{-1}\mathbf{C}.$$

4. Write

$$\mathbf{P}_2 = \mathbf{M}_2 - \mathbf{B}$$

and verify that **B** must equal

$$\mathbf{B} = \mathbf{M}_2\mathbf{C}^T(\mathbf{CM}_2\mathbf{C}^T + \mathbf{\Lambda_w})^{-1}\mathbf{CM}_2.$$

5. Verify that the answer to part 3 can be written as

$$\hat{\mathbf{a}}_2 = \mathbf{\Phi}\hat{\mathbf{a}}_1 + \mathbf{M}_2\mathbf{C}^T(\mathbf{CM}_2\mathbf{C}^T + \mathbf{\Lambda_w})^{-1}(\mathbf{R}_2 - \mathbf{C\Phi}\hat{\mathbf{a}}_1).$$

Compare the two forms with respect to ease of computation. What is the dimension of the matrix to be inverted?

Part III

1. Extend the results of Parts I and II to find an expression for $\hat{\mathbf{a}}_k$ and $\mathbf{P}_k$ in terms of $\hat{\mathbf{a}}_{k-1}$ and $\mathbf{M}_k$. The resulting equations are called the Kalman filter equations for discrete systems [38].
2. Draw a block diagram of the optimum processor.

Part IV. Verify that the Kalman filter reduces to the result in Problem 2.6.13 when $\mathbf{\Phi} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{0}$.

Special Applications

A large number of problems in the areas of pattern recognition, learning systems, and system equalization are mathematically equivalent to the general Gaussian problem. We consider three simple problems (due to M. E. Austin) in this section. Other examples more complex in detail but not in concept are contained in the various references.

Problem 2.6.16. Pattern Recognition. A pattern recognition system is to be implemented for the classification of noisy samples taken from a set of M patterns. Each pattern may be represented by a set of parameters in which the mth pattern is characterized by the vector $\mathbf{s}_m$. In general, the $\mathbf{s}_m$ vectors are unknown. The samples to be classified are of the form

$$\mathbf{x} = \mathbf{s}_m + \mathbf{n},$$

where the $\mathbf{s}_m$ are assumed to be independent Gaussian random variables with mean $\bar{\mathbf{s}}_m$ and covariance $\mathbf{\Lambda}_m$ and $\mathbf{n}$ is assumed to be zero-mean Gaussian with covariance $\mathbf{\Lambda_n}$ independent from sample to sample, and independent of $\mathbf{s}_m$.

1. In order to classify the patterns the recognition systems needs to know the pattern charac'eristics. We provide it with a "learning" sample:

$$\mathbf{x}_m = \mathbf{s}_m + \mathbf{n},$$

where the system knows that the mth pattern is present.

Show that if J learning samples, $\mathbf{x}_m^{(1)}, \mathbf{x}_m^{(2)}, \ldots, \mathbf{x}_m^{(J)}$, of the form $\mathbf{x}_m^{(j)} = \mathbf{s}_m + \mathbf{n}^{(j)}$ are available for each $m = 1, \ldots, M$, the pattern recognition system need store only the quantities

$$\mathbf{l}_m = \frac{1}{J} \sum_{j=1}^{J} \mathbf{x}_m{}^{(j)}$$

for use in classifying additional noisy samples; that is, show that the $\mathbf{l}_m$, $m = 1, \ldots M$' form a set of sufficient statistics extracted from the MJ learning samples.

2. What is the MAP estimate of $\mathbf{s}_m$? What is the covariance of this estimate as a function of J, the number of learning samples?

3. For the special case of two patterns ($M = 2$) characterized by unknown scalars s_1 and s_2, which have a priori densities $N(\bar{s}_1, \sigma)$ and $N(\bar{s}_2, \sigma)$, respectively, find the optimum decision rule for equiprobable patterns and observe that this approaches the decision rule of the "known patterns" classifier asymptotically with increasing number of learning samples J.

Problem 2.6.17. Intersymbol Interference. Data samples are to be transmitted over a known dispersive channel with an impulse response $h(t)$ in the presence of white Gaussian noise. The received waveform

$$r(t) = \sum_{k=-K}^{K} \xi_k h(t - kT) + n(t)$$

may be passed through a filter matched to the channel impulse response to give a set of numbers

$$a_j = \int r(t)\, h(t - jT)\, dt$$

for $j = 0, \pm 1, \pm 2, \ldots, \pm K$, which forms a set of sufficient statistics in the MAP estimation of the ξ_k. (This is proved in Chapter 4.) We denote the sampled channel autocorrelation function as

$$b_j = \int h(t)\, h(t - jT)\, dt$$

and the noise at the matched filter output as

$$n_j = \int n(t)\, h(t - jT)\, dt.$$

The problem then reduces to an estimation of the ξ_k, given a set of relations

$$a_j = \sum_{k=-K}^{K} \xi_k b_{j-k} + n_j \qquad \text{for } j, k = 0, \pm 1, \pm 2, \ldots \pm K.$$

Using obvious notation, we may write these equations as

$$\mathbf{a} = \mathbf{B}\boldsymbol{\xi} + \mathbf{n}.$$

1. Show that if $n(t)$ has double-sided spectral height $\frac{1}{2}N_0$, that the noise vector $\mathbf{n}$ has a covariance matrix $\mathbf{\Lambda_n} = \frac{1}{2}N_0\mathbf{B}$.

2. If the ξ_k are zero-mean Gaussian random variables with covariance matrix $\mathbf{\Lambda}_{\boldsymbol{\xi}}$ show that the MAP estimate of $\boldsymbol{\xi}$ is of the form $\boldsymbol{\xi} = \mathbf{Ga}$ and therefore that $\hat{\boldsymbol{\xi}}_0 = \mathbf{g}^T\mathbf{a}$. Find $\mathbf{g}$ and note that the estimate of ξ_0 can be obtained by passing the sufficient statistics into a tapped delay line with tap gains equal to the elements of $\mathbf{g}$. This cascading of a matched filter followed by a sampler and a transversal filter is a well-known equalization method employed to reduce intersymbol interference in digital communication via dispersive media.

Problem 2.6.18. Determine the MAP estimate of ξ_0 in Problem 2.6.17; assuming further that the ξ_k are independent and that the ξ_k are known (say through a "teacher" or infallible estimation process) for $k < 0$. Show then that the weighting of the sufficient statistics is of the form

$$\hat{\xi}_0 = \sum_{j>0} g_j a_j - \sum_{j<0} f_j \xi_j$$

and find g_j and f_j. This receiver may be interpreted as passing the sampled matched-filter output through a transversal filter with tap gains g_j and subtracting the output from a second transversal filter whose input is the sequence of ξ_k which estimates have been made. Of course, in implementation such a receiver would be self-taught by using its earlier estimates as correct in the above estimation equation.

Problem No. 2.6.19. Let

$$z = \mathbf{G}^T\mathbf{r}$$

and assume that z is $N(m_z, \sigma_z)$ for all finite $\mathbf{G}$.

1. What is $M_z(jv)$? Express your result in terms of m_z and σ_z.
2. Rewrite the result in (1) in terms of $\mathbf{G}$, $\mathbf{m}$, and $\mathbf{\Lambda_r}$ [see (316)–(317) for definitions].
3. Observe that

$$M_z(ju) \triangleq E[e^{juz}] = E[e^{ju\mathbf{G}^T\mathbf{r}}]$$

and

$$M_{\mathbf{r}}(j\mathbf{v}) \triangleq E[e^{j\mathbf{v}^T\mathbf{r}}]$$

and therefore

$$M_z(ju) = M_{\mathbf{r}}(j\mathbf{v}) \quad \text{if} \quad \mathbf{G}u = \mathbf{v}.$$

Use these observations to verify (317).

Problem No. 2.6.20 (***continuation***).

(a) Assume that the $\mathbf{\Lambda_r}$ defined in (316) is positive definite. Verify that the expression for $p_{\mathbf{r}}(\mathbf{R})$ in (318) is correct. [*Hint*. Use the diagonalizing transformation $\mathbf{W}$ defined in (368).]

(b) How must (318) be modified if $\mathbf{\Lambda_r}$ is singular? What does this singularity imply about the components of $\mathbf{r}$?

P2.7 Performance Bounds and Approximations

Problem 2.7.1. Consider the binary test with N independent observations, r_i, where

$$p_{r_i|H_k} = N(m_k, \sigma_k), \qquad \begin{aligned} k &= 0, 1, \\ i &= 1, 2, \dots, N. \end{aligned}$$

Find $\mu(s)$.

Problem 2.7.2 (***continuation***). Consider the special case of Problem 2.7.1 in which

$$m_0 = 0,$$
$$\sigma_0^2 = \sigma_n^2,$$

and

$$\sigma_1^2 = \sigma_s^2 + \sigma_n^2.$$

1. Find $\mu(s)$, $\dot{\mu}(s)$, and $\ddot{\mu}(s)$.
2. Assuming equally likely hypotheses, find an upper bound on the minimum $\Pr(\epsilon)$.
3. With the assumption in part 2, find an approximate expression for the $\Pr(\epsilon)$ that is valid for large N.

Problem 2.7.3. A special case of the binary Gaussian problem with **N** observations is

$$p_{\mathbf{r}|H_k}(\mathbf{R}|H_k) = \frac{1}{(2\pi)^{N/2}|\mathbf{K}_k|^{\frac{1}{2}}} \exp\left(-\frac{\mathbf{R}^T\mathbf{K}_k^{-1}\mathbf{R}}{2}\right), \qquad k = 0, 1.$$

1. Find $\mu(s)$.
2. Express it in terms of the eigenvalues of the appropriate matrices.

Problem 2.7.4 (***continuation***). Consider the special case in which

$$\mathbf{K}_0 = \sigma_n^2\mathbf{I}$$

and

$$\mathbf{K}_1 = \mathbf{K}_s + \mathbf{K}_0.$$

Find $\mu(s)$, $\dot{\mu}(s)$, $\ddot{\mu}(s)$.

Problem 2.7.5 (***alternate continuation of 2.7.3***). Consider the special case in which $\mathbf{K}_1$ and $\mathbf{K}_0$ are partitioned into the 4 $N \times N$ matrices given by (422) and (423).

1. Find $\mu(s)$.
2. Assume that the hypotheses are equally likely and that the criterion is minimum $\Pr(\epsilon)$. Find a bound on the $\Pr(\epsilon)$.
3. Find an approximate expression for the $\Pr(\epsilon)$.

Problem 2.7.6. The general binary Gaussian problem for N observations is

$$p_{\mathbf{r}|H_k}(\mathbf{R}|H_k) = \frac{1}{(2\pi)^{N/2}|\mathbf{K}_k|^{\frac{1}{2}}} \exp\left[-\frac{(\mathbf{R}^T - \mathbf{m}_k^T)\mathbf{K}_k^{-1}(\mathbf{R} - \mathbf{m}_k)}{2}\right], \qquad k = 0, 1.$$

Find $\mu(s)$.

Problem 2.7.7. Consider Example 3A on p. 130. A bound on the $\Pr(\epsilon)$ is

$$\Pr(\epsilon) \le \tfrac{1}{2}\left[\frac{(1 + \sigma_s^2/\sigma_n^2)}{(1 + \sigma_s^2/2\sigma_n^2)^2}\right]^{N/2}$$

1. Constrain $N\sigma_s^2/\sigma_n^2 = x$. Find the value of N that minimizes the bound.
2. Evaluate the approximate expression in (516) for this value of N.

Problem 2.7.8. We derived the Chernoff bound in (461) by using tilted densities. This approach prepared us for the central limit theorem argument in the second part of our discussion. If we are interested only in (461), a much simpler derivation is possible.

1. Consider a function of the random variable x which we denote as $f(x)$. Assume

$$f(x) \ge 0, \qquad \text{all } x,$$
$$f(x) \ge f(X_0) > 0, \qquad \text{all } x \ge X_0.$$

Prove

$$\Pr[x \ge X_0] \le \frac{E[f(x)]}{f(X_0)}.$$

2. Now let

$$f(x) = e^{sx}, \qquad s \ge 0,$$

and

$$X_0 = \gamma.$$

Use the result in (1) to derive (457). What restrictions on γ are needed to obtain (461)?

Problem 2.7.9. The reason for using tilted densities and Chernoff bounds is that a straightforward application of the central limit theorem gives misleading results when the region of interest is on the tail of the density. A trivial example taken from [4-18] illustrates this point.

Consider a set of statistically independent random variables x_i which assumes values 0 and 1 with equal probability. We are interested in the probability

$$\Pr\left[y_N = \frac{1}{N}\sum_{i=1}^{N} x_i \ge 1\right] \triangleq \Pr[A_N].$$

(a) Define a standardized variable

$$z \triangleq \frac{y_N - \bar{y}_N.}{\sigma_{y_N}}$$

Use a central limit theorem argument to estimate $\Pr[A_N]$. Denote this estimate as $\hat{\Pr}[A_N]$.

(b) Calculate $\Pr[A_N]$ exactly.

(c) Verify that the fractional error is,

$$\frac{\hat{\Pr}[A_N]}{\Pr[A_N]} \propto e^{0.19N}$$

Observe that the fractional error grows exponentially with N.

(d) Estimate $\Pr[A_N]$ using the Chernoff bound of Problem 2.7.8. Denote this estimate as $\Pr_c[A_N]$. Compute $\dfrac{\Pr_c[A_N]}{\Pr[A_N]}$.

REFERENCES

[1] W. B. Davenport and W. L. Root, *Random Signals and Noise*, McGraw-Hill, New York, 1958.

[2] A. T. Bharucha-Reid, *Elements of the Theory of Markov Processes and their Applications*, McGraw-Hill, New York, 1960.

[3] M. Abramovitz and I. A. Stegun (ed.), *Handbook of Mathematical Functions*, National Bureau of Standards, U.S. Government Printing Office, Washington, D.C., June 1964.

[4] J. A. Greenwood and H. O. Hartley, *Guides to Tables in Mathematical Statistics*, Princeton University Press, Princeton, New Jersey, 1962.

[5] R. A. Fisher, "Theory of Statistical Estimation," *Proc. Cambridge Phil. Soc.*, **22**, 700 (1925).

[6] R. A. Fisher, "On the Mathematical Foundations of Theoretical Statistics," *Phil. Trans. Roy. Soc., London*, **222**, 309 (1922).

[7] R. A. Fisher, "Two New Properties of Mathematical Likelihood," *Proc. Roy. Soc., London*, **144**, 285 (1934).

[8] R. A. Fisher, "The Logic of Inductive Inference," *J. Roy. Statist. Soc.*, **98**, 39 (1935).

[9] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, New Jersey, 1946.

[10] S. S. Wilks, *Mathematical Statistics*, Wiley, New York, 1962.

[11] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics, Vol. 2, Inference and Relationship*, Hafner, New York, 1961.

[12] C. R. Rao, "Information and Accuracy Attainable in the Estimation of Statistical Parameters," *Bull. Calcutta Math. Soc.*, **37**, 81–91 (1945).

[13] A. Bhattacharyya, "On Some Analogues of the Amount of Information and their Use in Statistical Estimation," *Sankhya*, **8**, 1, 201, 315 (1946, 1947, 1948).

[14] H. L. Van Trees, "A Generalized Bhattacharyya Bound," Internal Memo., Detection & Estimation Theory Group, MIT, 1966.

[15] E. W. Barankin, "Locally Best Unbiased Estimates," *Ann. Math. Stat.*, **20**, 477 (1949).

[16] R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

[17] E. L. Lehmann, *Testing Statistical Hypotheses*, Wiley, New York, 1959.

[18] F. B. Hildebrand, *Methods of Applied Mathematics*, Prentice-Hall, Englewood Cliffs, New Jersey, 1952.

[19] R. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, Oliver & Boyd, Edinburgh, 1953.

[20] S. Sherman, "Non-mean-square Error Criteria," *IRE Trans. Information Theory*, **IT-4**, No. 3, 125–126 (1958).

[21] K. Pearson, *Tables of the Incomplete* Γ*-Function*, Cambridge University Press, Cambridge, 1934.

[22] J. N. Pierce, "Theoretical Diversity Improvement in Frequency-Shift Keying," *Proc. IRE*, **46**, 903–910 (May 1958).

[23] C. E. Shannon, "Seminar Notes for Seminar in Information Theory," MIT, 1956 (unpublished).

[24] R. M. Fano, *Transmission of Information*, MIT Press, Cambridge, Massachusetts, and Wiley, New York, 1961.

[25] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels: I," *Information and Control*, Vol. **10**, No. 1, 65–103 (1967).

[26] R. G. Gallager, "Lower Bounds on the Tails of Probability Distributions," MIT, RLE, QPR 77, 277–291 (April 1965).

[27] I. M. Jacobs, "Probability-of-Error Bounds for Binary Transmission on the Slow Fading Rician Channel," *IEEE Trans. Information Theory*, Vol. **IT-12**, No. 4, Oct, 1966.

[28] H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of a Hypothesi based on the Sum of Observations," *Annals Math. Stat.*, **23**, 493–507 (1962).

[29] A. Bhattacharyya, "On a Measure of Divergence Between Two Statistical Populations defined by their Probability Distributions," *Bull. Calcutta Math. Soc.*, **35**, No. 3, 99–110 (1943).

[30] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, Wiley, New York, 1950, 1957.

[31] D. Dugué, "Application des Proprietes de la Limite au Sens du Calcul des Probabilities a L'etude des Diverses Questions D'estimation," *Ecol. Poly.*, **3**, No. 4, 305–372 (1937).

[32] A. Sommerfeld, *An Introduction to the Geometry of N Dimensions*, Dutton, New York, 1929.

[33] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, Wiley, New York, 1966.

[34] H. L. Yudkin, "An Error Bound for Gaussian Signals in Gaussian Noise," MIT, RLE, QPR 73, April 15, 1964.

[35] T. J. Goblick, "Study of An Orbiting Dipole Belt Communication System," Lincoln Laboratory, Technical Report 369, December 22, 1964.

[36] A. J. Viterbi, *Principles of Coherent Communication*, McGraw-Hill, New York, 1966.

[37] K. Pearson, "On the Systematic Fitting of Curves to Observations and Measurements," *Biometrika*, **1**, 265 (1902).

[38] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *J. Basic Eng.* (ASME Trans.) **82D**, 35–45 (1960).