Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory. Harry L. Van Trees

ISBNs: 0-471-09517-6 (Paperback); 0-471-22108-2 (Electronic)

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

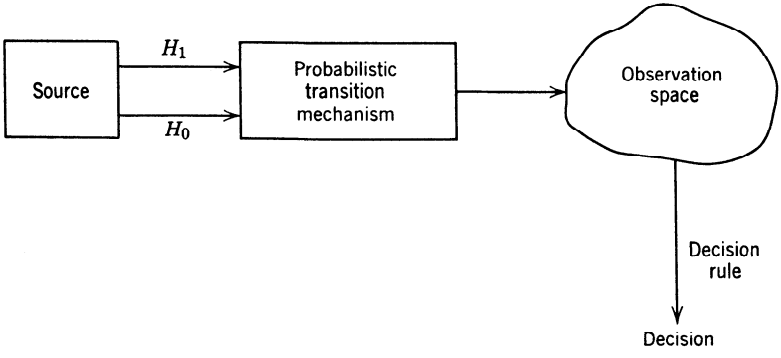


Fig. 2.1 Components of a decision theory problem.

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

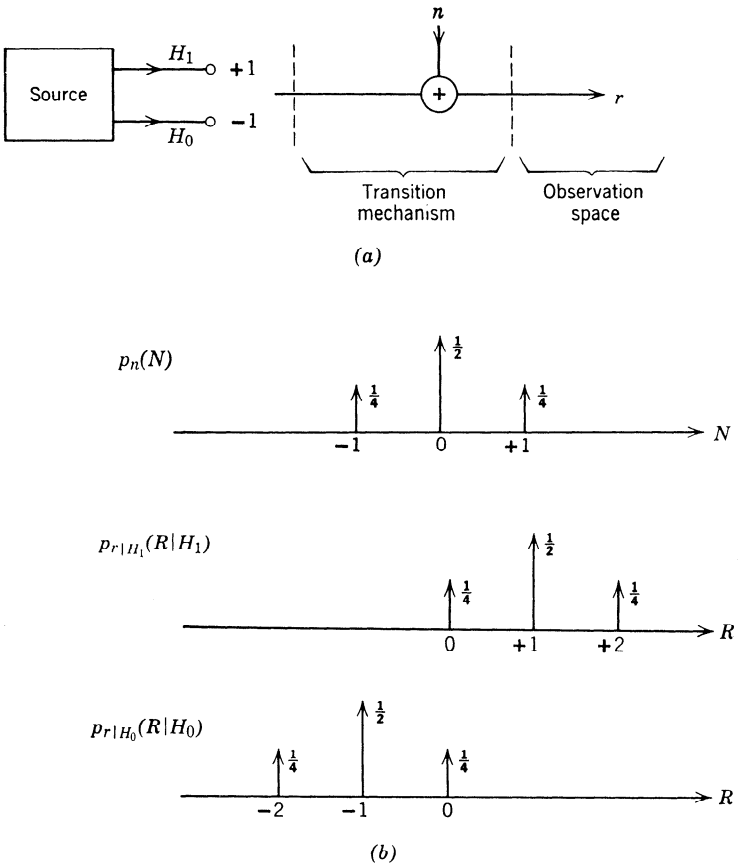


Fig. 2.2 A simple decision problem: (*a*) model; (*b*) probability densities.

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.

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

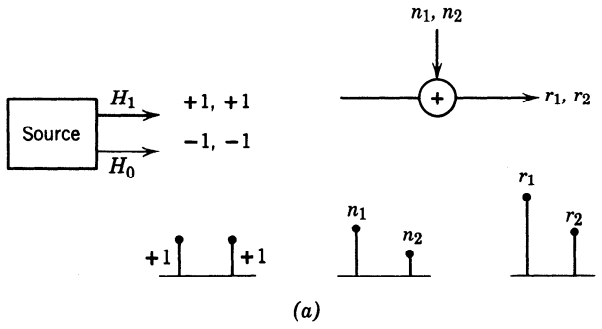
$$\begin{aligned} H_1&: r_1 = 1 + n_1 \\ &\quad r_2 = 1 + n_2, \\ H_0&: r_1 = -1 + n_1 \\ &\quad r_2 = -1 + n_2. \end{aligned} \tag{2}$$

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.



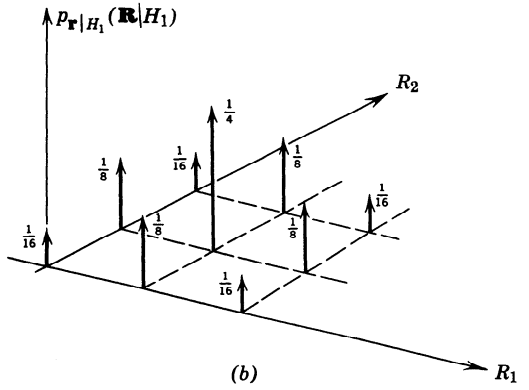


Fig. 2.3 A two-dimensional problem: (*a*) model; (*b*) probability density.

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.

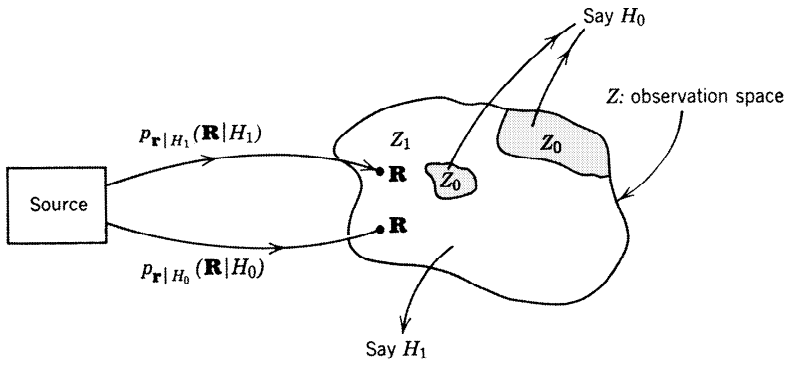


Fig. 2.4 Decision regions.

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

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

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

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

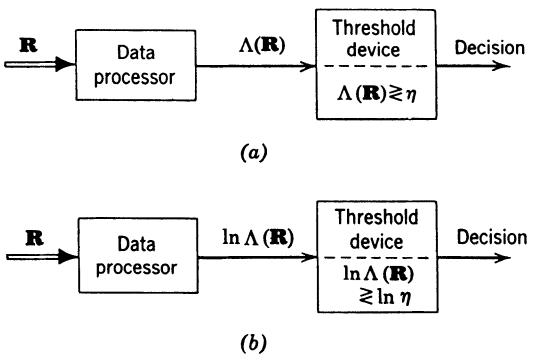


Fig. 2.5 Likelihood ratio processors.

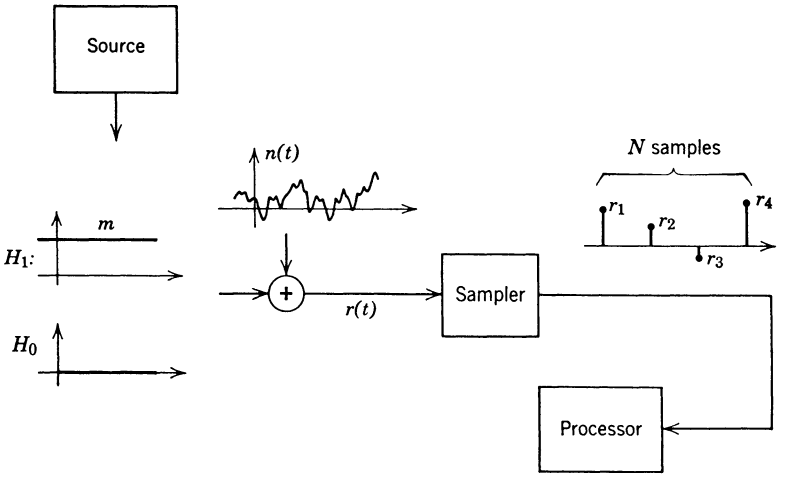


Fig. 2.6 Model for Example 1.

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

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

and

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

Substituting into (13), we have

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

After canceling common terms and taking the logarithm, we have

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

Thus the likelihood ratio test is

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

or, equivalently,

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

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

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

and

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

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

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

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

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

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

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

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

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

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,

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

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

$$E(n) = m_i. \tag{34}$$

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

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

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 integrals are determined. We denote these values in the following manner:

$$P_F = \int_{Z_1} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R},$$

$$P_D = \int_{Z_1} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R}, \tag{41}$$

$$P_M = \int_{Z_0} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R} = 1 - P_D.$$

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):

$$\mathcal{R} = P_0C_{10} + P_1C_{11} + P_1(C_{01} - C_{11})P_M - P_0(C_{10} - C_{00})(1 - P_F). \tag{42}$$

Because

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

(42) becomes

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

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,

$$\mathcal{R}_F(P_1^*, P_1) \geq \mathcal{R}_B(P_1). \tag{45}$$

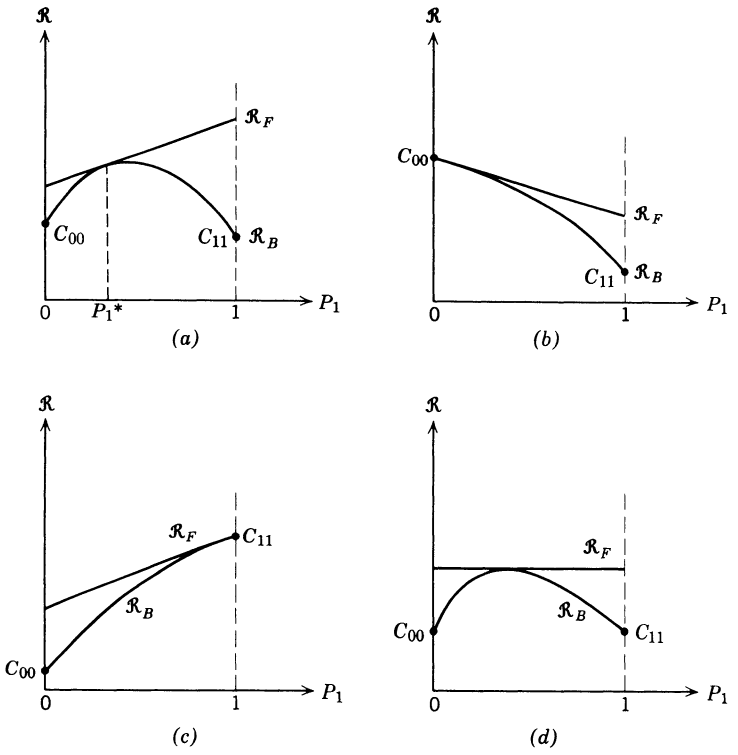


Fig. 2.7 Risk curves: (*a*) fixed risk versus typical Bayes risk; (*b*) maximum value of $\mathcal{R}_1$ at $P_1 = 0$.

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 $P_F = \alpha' \leq \alpha$ 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 $\alpha' \leq \alpha$. 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.

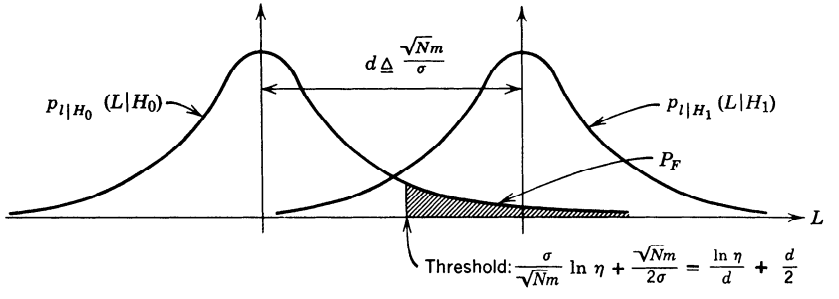
2.2.2 Performance: Receiver Operating Characteristic

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.

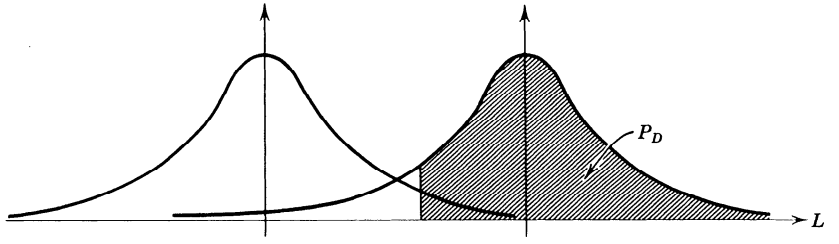
We begin by considering Example 1 in Section 2.2.1.

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

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



(*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

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

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

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

is its complement. In this notation

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

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*:

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

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 $P_F = P_D = 0$.

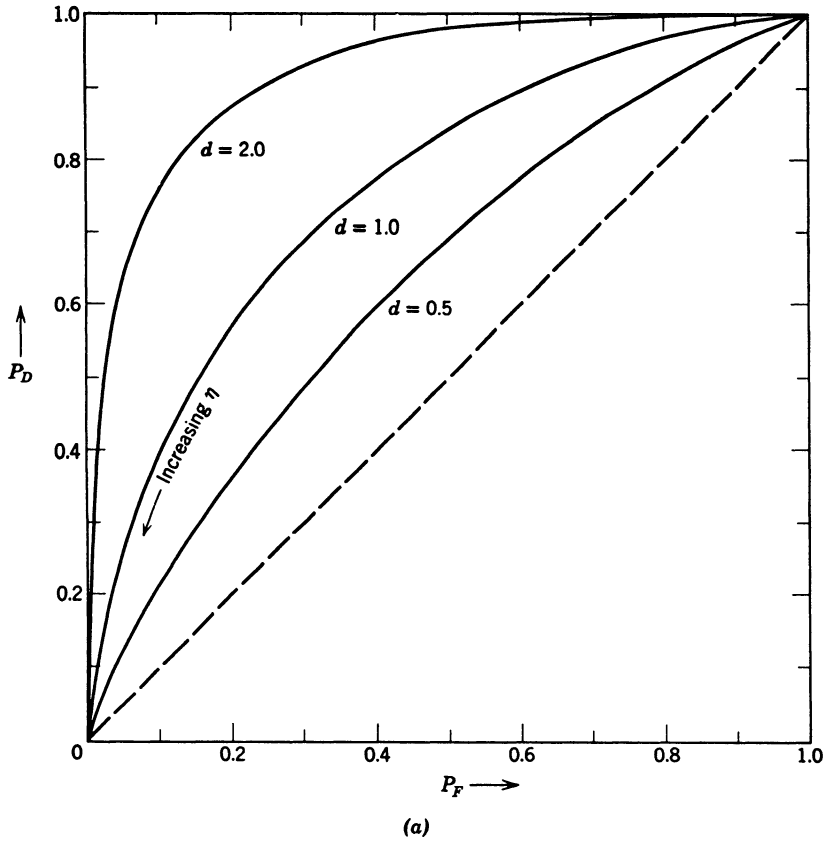
As we would expect from Fig. 2.8, the performance increases monotonically with d. In Fig. 2.9*b* we have replotted the results to give P_D versus d with P_F as a parameter on the curves. For a particular d we can obtain any point on the curve by choosing η appropriately ($0 \leq \eta \leq \infty$).

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 $\text{erf}(X) = \sqrt{2/\pi} \int_0^X \exp(-y^2)\,dy$, 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.

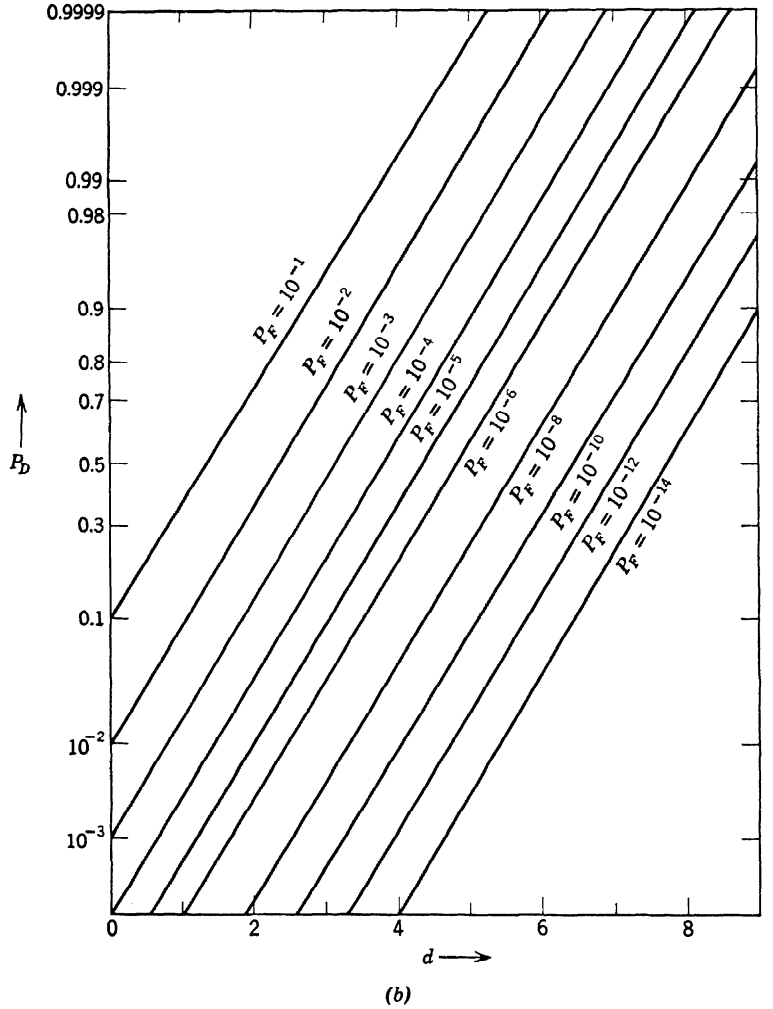
Before calculating the performance of the other two examples, it is worthwhile to point out two simple bounds on $\operatorname{erfc}_*(X)$. They will enable

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

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

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

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



(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.

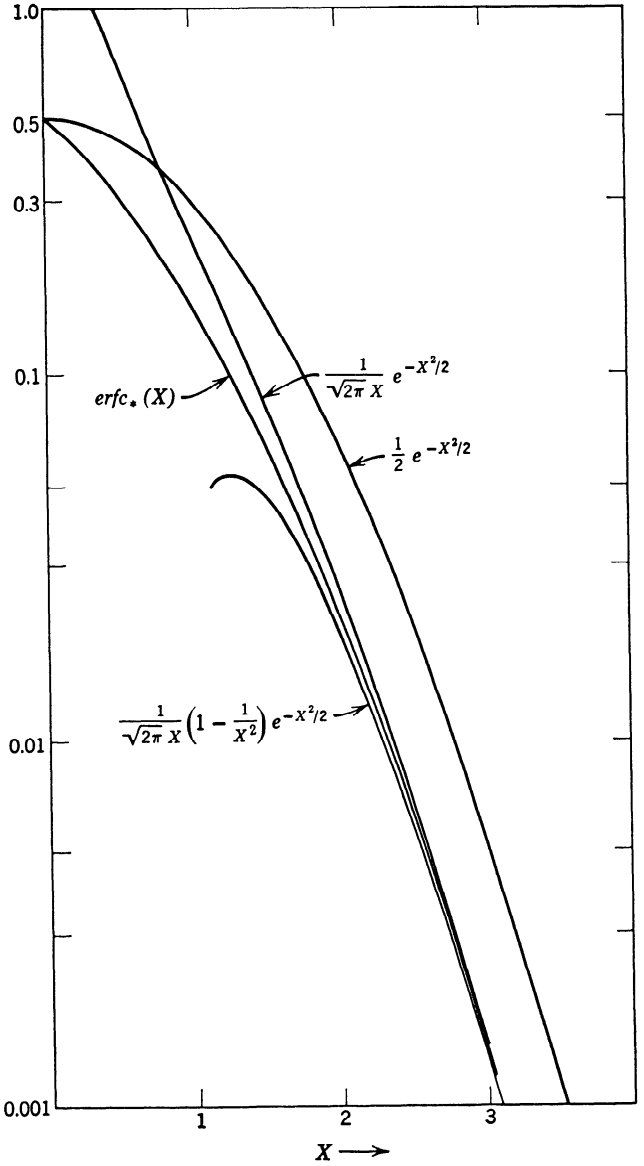


Fig. 2.10 Plot of $\text{erfc}_*(X)$ and related functions.

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

$$n \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_I, \qquad \gamma_I = 0, 1, 2, \ldots, \tag{83}$$

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	$1 - e^{-m_0}$	$1 - e^{-m_1}$

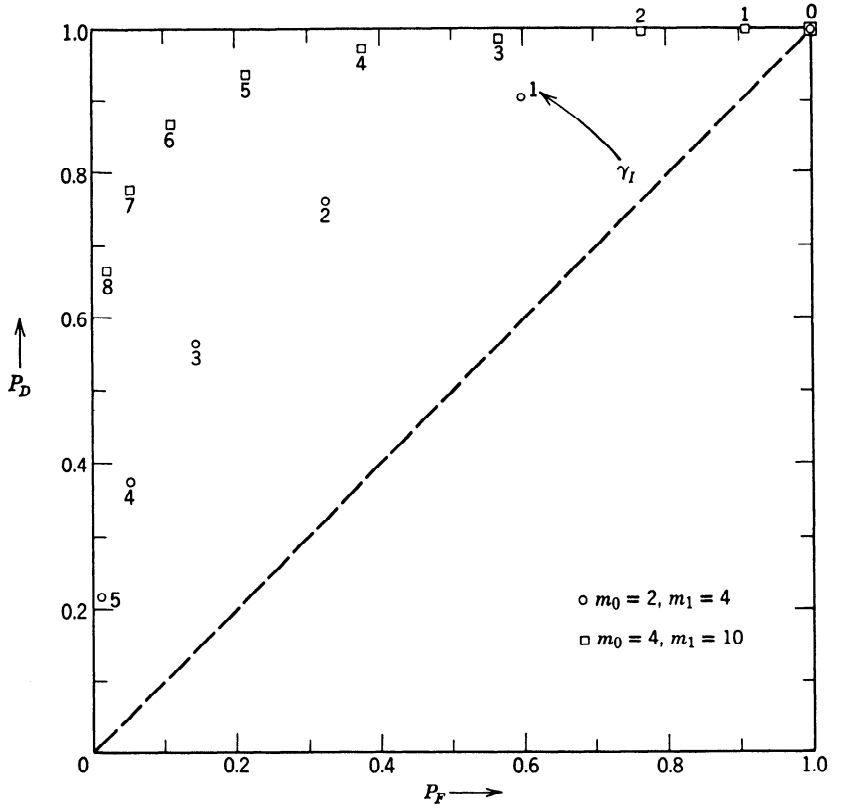


Fig. 2.11 (*a*) **Receiver operating characteristic, Poisson problem.**

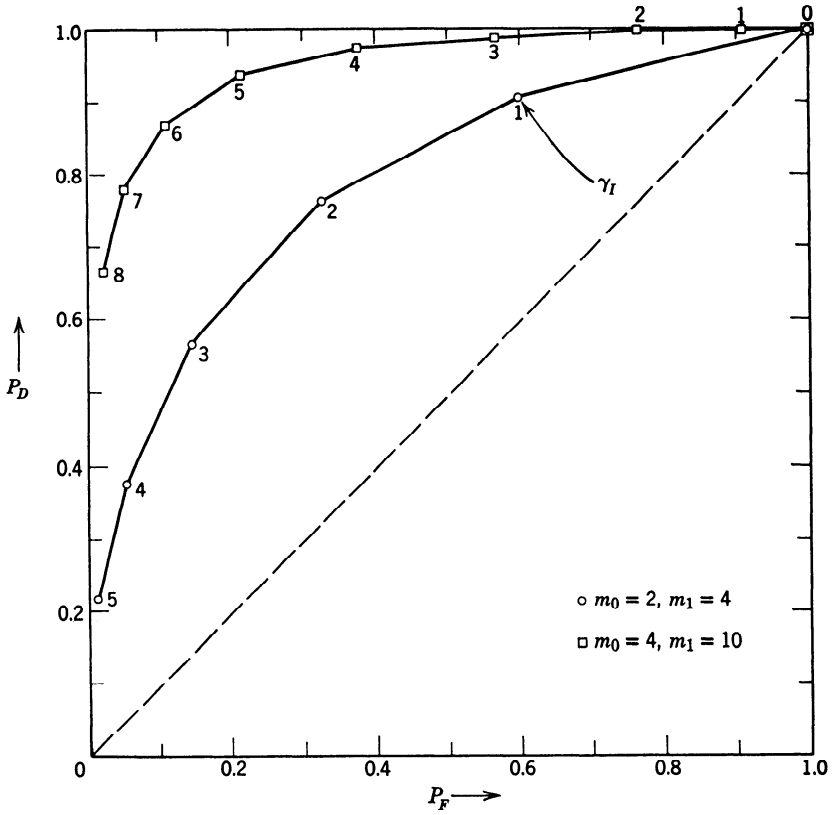


Fig. 2.11 (*b*) **Receiver operating characteristic with randomized decision rule.**

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

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

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

$$P_D(\eta) \triangleq \Pr\{\Lambda(\mathbf{R}) \geq \eta|H_1\} = \int_{\Omega(\eta)} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R}$$

$$= \int_{\Omega(\eta)} \Lambda(\mathbf{R}) p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R}, \tag{92}$$

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

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

Differentiating (93) with respect to η, we obtain

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

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

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

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

$$C_F P_F = C_M P_M = C_M(1 - P_D), \quad (96)$$

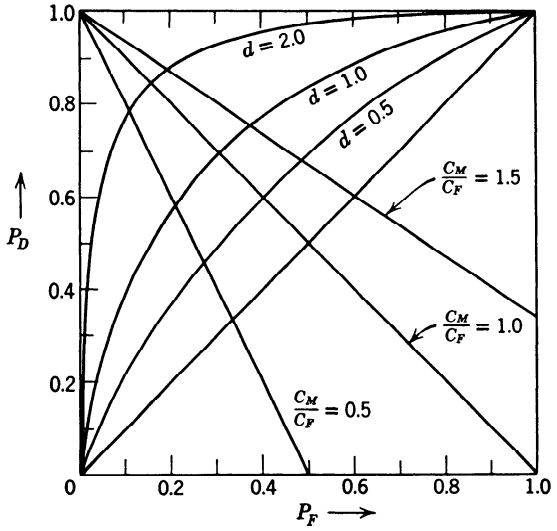


Fig. 2.12 Determination of minimax operating point.

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

$$\begin{aligned}
\mathcal{R} = {} & P_0 C_{00} \int_{Z-Z_1-Z_2} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} + P_0 C_{10} \int_{Z_1} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} \\
& + P_0 C_{20} \int_{Z_2} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} + P_1 C_{11} \int_{Z-Z_0-Z_2} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R} \\
& + P_1 C_{01} \int_{Z_0} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R} + P_1 C_{21} \int_{Z_2} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R} \\
& + P_2 C_{22} \int_{Z-Z_0-Z_1} p_{\mathbf{r}|H_2}(\mathbf{R}|H_2)\, d\mathbf{R} + P_2 C_{02} \int_{Z_0} p_{\mathbf{r}|H_2}(\mathbf{R}|H_2)\, d\mathbf{R} \\
& + P_2 C_{12} \int_{Z_1} p_{\mathbf{r}|H_2}(\mathbf{R}|H_2)\, d\mathbf{R}.
\end{aligned} \tag{99}$$

This reduces to

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

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:

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

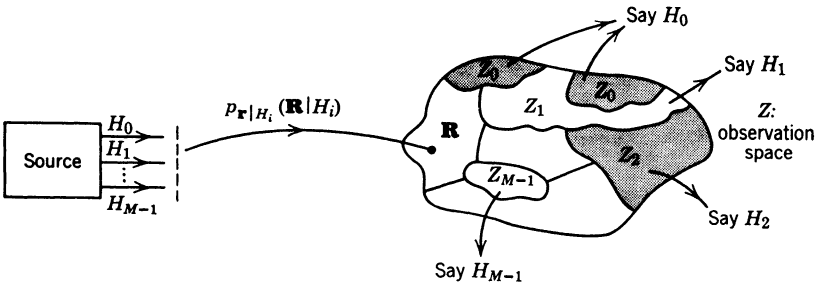


Fig. 2.13 *M* **hypothesis problem.**

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.

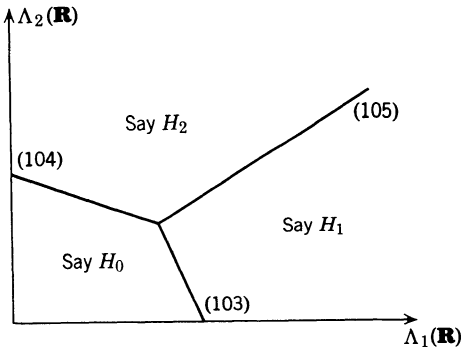
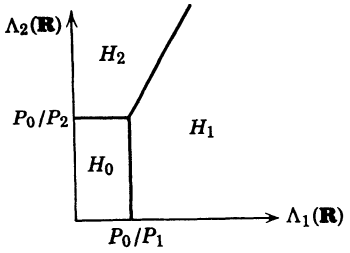


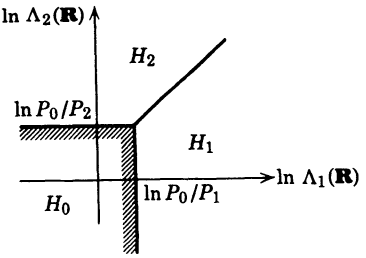
Fig. 2.14 Decision space.

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

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

and (105) becomes an identity.

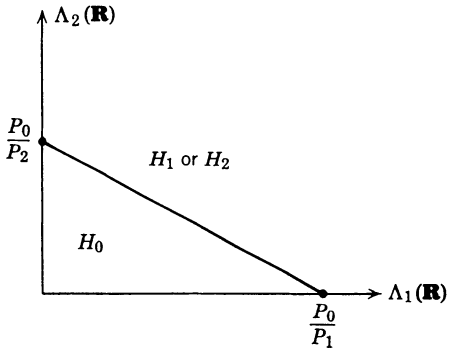


Fig. 2.16 Decision spaces.

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

$$P_0 = 0, \qquad P_1 + P_2 = 1,$$

and

$$C_{12} = C_{02}, C_{21} = C_{01}. \tag{114}$$

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 examples

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

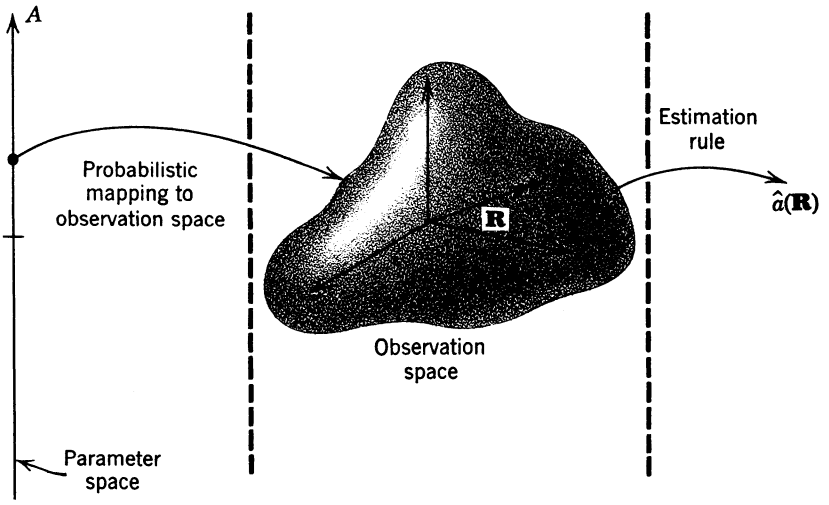


Fig. 2.17 Estimation model.

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 $\pm\Delta/2$. In other words, an error less than $\Delta/2$ in magnitude is as good as no error. If $a_\epsilon > \Delta/2$, we assign a uniform value:

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

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

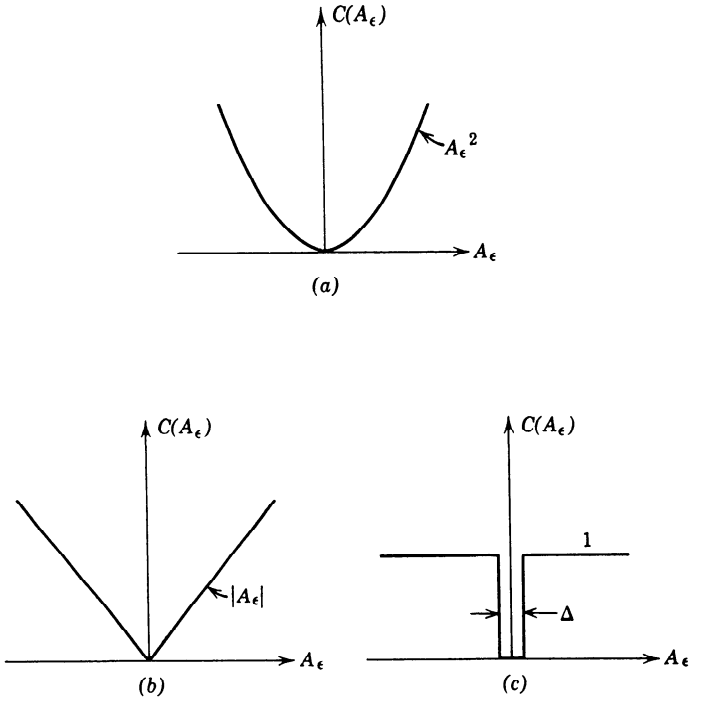


Fig. 2.18 Typical cost functions: (*a*) mean-square error; (*b*) absolute error; (*c*) uniform cost function.

problems of interest the same estimate will be optimum for a large class of cost functions.

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:

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

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

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

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 $\mathcal{R}_{ms}$. Substituting (119) into (123), we have

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

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

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

Now the inner integral and $p_{\mathbf{r}}(\mathbf{R})$ are non-negative. Therefore we can minimize $\mathcal{R}_{ms}$ by minimizing the inner integral. We denote this estimate $\hat{a}_{ms}(\mathbf{R})$. 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

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

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 $\mathcal{R}_{ms}$ 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

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

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

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

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:

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

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 $\hat{a}_{map}(\mathbf{R})$, the *maximum a posteriori* estimate. In the sequel we use $\hat{a}_{map}(\mathbf{R})$ without further reference to the uniform cost function.

To find $\hat{a}_{map}$ 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:

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

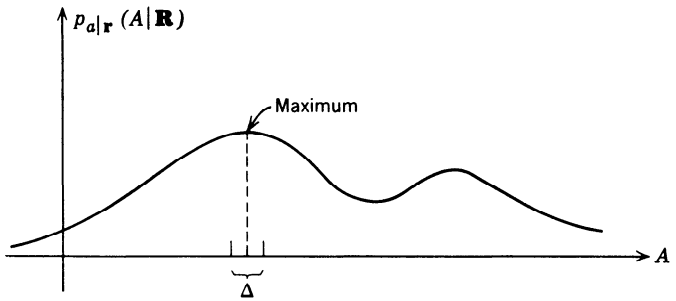


Fig. 2.19 An a posteriori density.

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

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

To find $\hat{a}_{\text{ms}}(\mathbf{R})$ 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

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

the symmetry of the a posteriori density implies

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

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

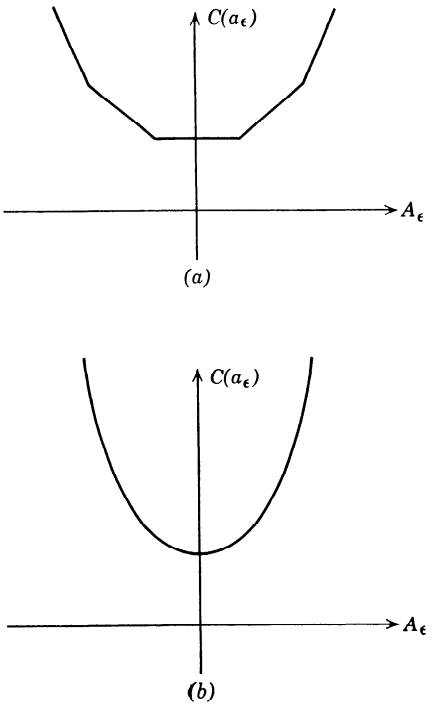


Fig. 2.20 Symmetric convex cost functions: (*a*) convex; (*b*) strictly convex.

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

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

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

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

[Use (151) in (153)]

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

[(152) implies this equality]

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

[(149) implies this equality]

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

[(152) implies this equality].

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

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

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)

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

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:

$$\hat{a}_{ms}(N) = \frac{(1+\lambda)^{N+1}}{N!}\int_0^\infty A^{N+1}\exp[-A(1+\lambda)]\,dA$$

$$= \frac{(1+\lambda)^{N+1}}{(1+\lambda)^{N+2}}(N+1) = \left(\frac{1}{\lambda+1}\right)(N+1). \tag{167}$$

To find $\hat{a}_{map}$ 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

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

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

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 of 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., Cramer [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,

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

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

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

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

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

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:

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

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.

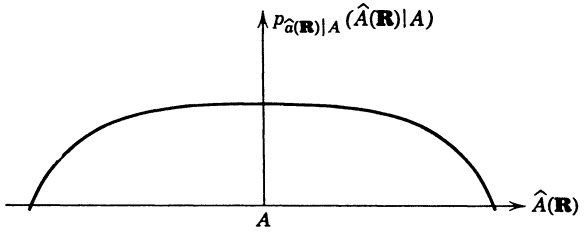


Fig. 2.21 Probability density for an estimate.

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

$$r = A + n, \tag{174}$$

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

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.

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

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:

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

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

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

or, equivalently,

(b)

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

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

$$\frac{\partial \ln \Pr(n = N|A)}{\partial A} = \frac{\partial}{\partial A}(N \ln A - A - \ln N!)$$

$$= \frac{N}{A} - 1 = \frac{1}{A}(N - A). \tag{202}$$

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

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

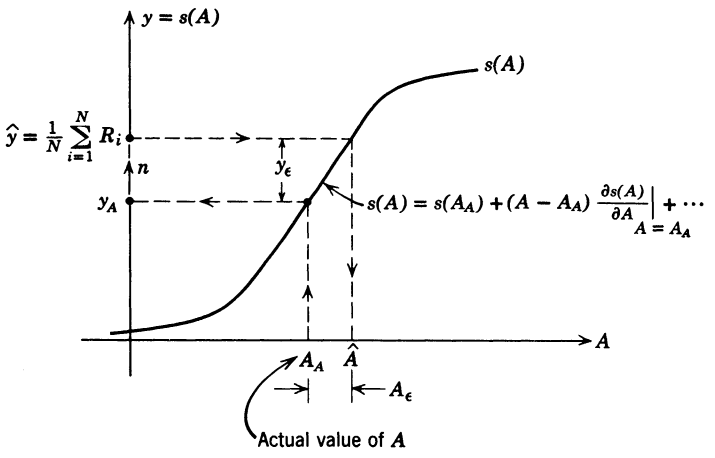


Fig. 2.22 Behavior of error variance in the presence of small errors.

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,

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

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

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

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:

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

Now integrate with respect to A:

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

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

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

or, equivalently,

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

with equality if and only if

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

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

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

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

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

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

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

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 $\hat{a}_{ms}(\mathbf{R}) = \hat{a}_{map}(\mathbf{R})$ 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 $\hat{a}_{ms}(\mathbf{R})$ or $\hat{a}_{map}(\mathbf{R})$, 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.

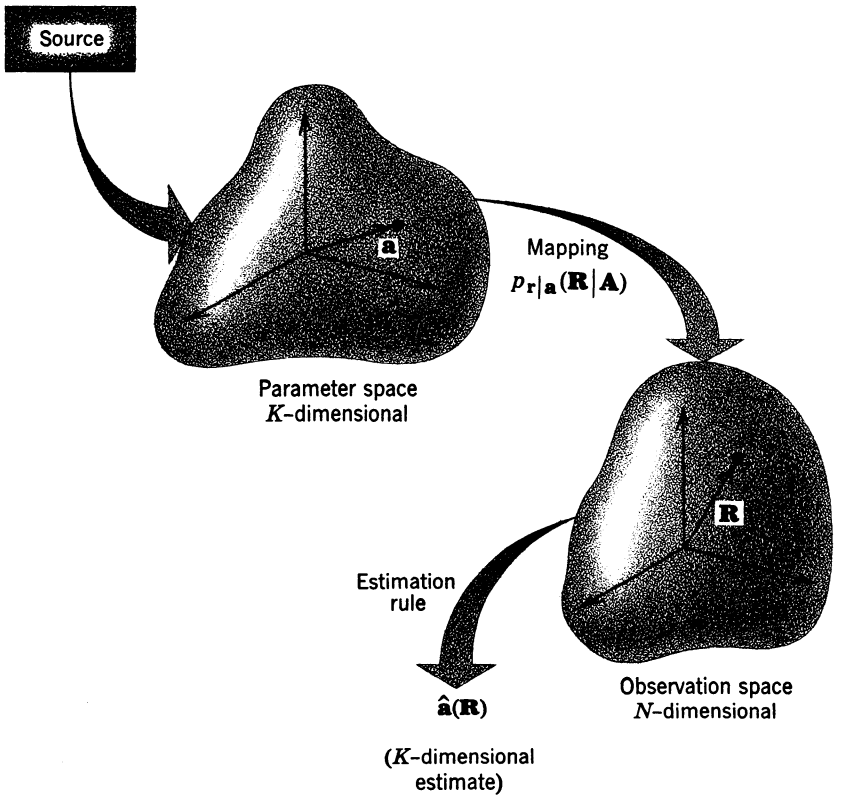


Fig. 2.23 Multiple parameter estimation model.

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

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

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 $\mathbf{A}$ 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:

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

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,

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

Similarly, for ML estimates we must find the value of $\mathbf{A}$ 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:

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

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

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

(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

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

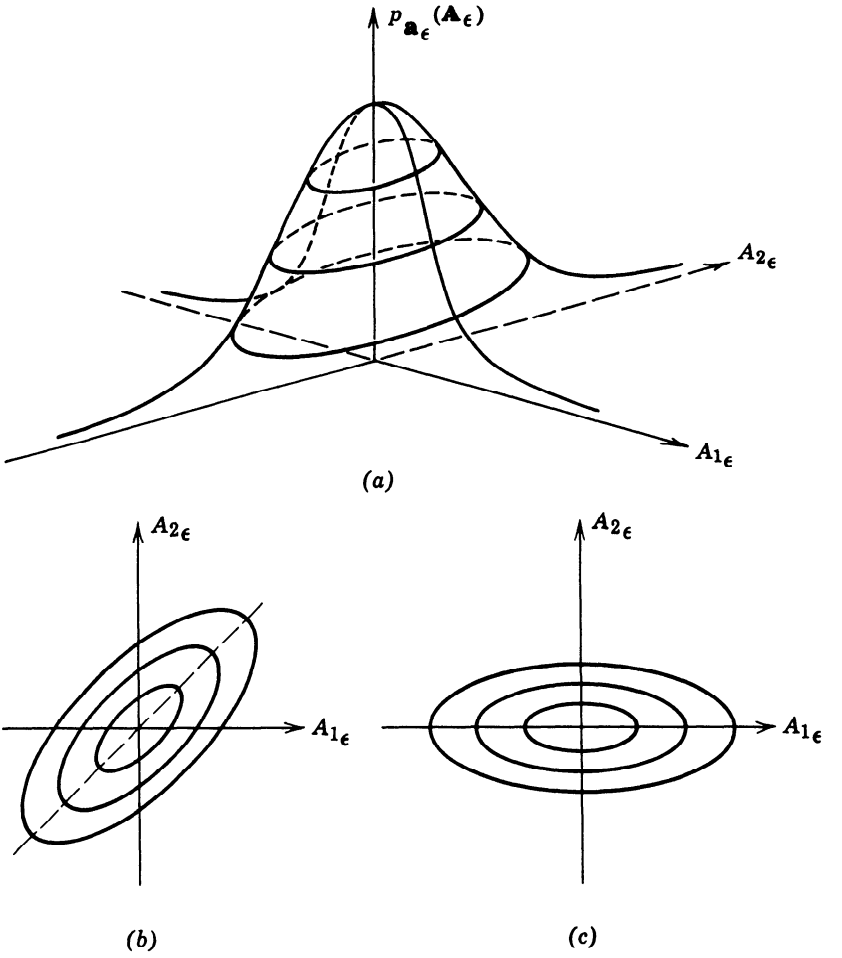


Fig. 2.24 Gaussian densities: [*a*] two-dimensional Gaussian density; [*b*] equal-height contours, correlated variables; [*c*] equal-height contours, uncorrelated variables.

The height of the probability density in this differential area is

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

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

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

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 $\mathbf{J}$ 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 $\mathbf{J}$ 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 $\mathbf{A}$. 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

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

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.

$$\text{Var}\,(d_{\epsilon_i}) \geq \sum_i^K \sum_j^K \frac{\partial g_{d_i}(\mathbf{A})}{\partial A_i} J^{ij} \frac{\partial g_{d_i}(\mathbf{A})}{\partial A_j}. \tag{287}$$

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 $\hat{\mathbf{d}}$ 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

$$\begin{aligned} J_{P_{ij}} &\triangleq E\left[\frac{\partial \ln p_{\mathbf{a}}(\mathbf{A})}{\partial A_i}\frac{\partial \ln p_{\mathbf{a}}(\mathbf{A})}{\partial A_j}\right] \\ &= -E\left[\frac{\partial^2 \ln p_{\mathbf{a}}(\mathbf{A})}{\partial A_i\,\partial A_j}\right]. \end{aligned} \tag{290}$$

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 if $\mathbf{J}_T$ 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

$$\begin{aligned} H_0 &: p_{r|H_0}(R|H_0) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{R^2}{2\sigma^2}\right), \\ H_1 &: p_{r|H_1,m}(R|H_1,M) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(R-M)^2}{2\sigma^2}\right), \qquad M_0 \le M \le M_1. \end{aligned} \tag{295}$$

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.