distributions is presumed. In contradistinction, Chapters 13 and 14 address the important case of linear estimators that require no knowledge of the distributions, i.e., they provide estimates that are distribution free. Section 10.2 presents a general formulation of parameter estimation where the observables are used to produce an estimate of the parameters of interest.

An important question to resolve is: How good is the estimate? Before addressing this question, it is useful to identify the relationship between estimation and detection theory, since many of the same concepts are used but from different viewpoints. The types of estimation problems are then summarized in Section 10.4 where two distinct classes are identified depending on whether or not distributions associated with the observables are known.

Returning to the question of the quality of the estimate, it is important to understand properties associated with a good estimate. Such properties include unbiased, consistent, invariant, minimum variance, and efficient. Section 10.5.4 describes estimators based on sufficient statistics, where the statistic summarizes the information about a parameter in a more manageable form and leads to a variance that is smaller than any other statistic not based on a sufficient statistic. Section 10.5.5 presents the important concept of the Cramér-Rao inequality. This inequality provides a bound on the minimum variance of the estimate. The minimum variance is shown to be determined from the inverse of the Fisher information, where a larger value of the Fisher information implies a smaller variance. If the Cramér-Rao bound is satisfied with equality, the estimate is seen to be most efficient. Extensions of the Cramér-Rao bound are addressed in Section 10.5.8 where the parameter may be a deterministic vector or a random scalar or vector. Specific estimation methods are then described depending on the knowledge of the distributions.

Bayes estimation, described in Section 10.6, is based on knowledge of the *a posteriori* distribution and a cost function, similar in function to Bayes detection. Bayes estimation is the most general method but requires the most information to be effectively used. Maximum *a posteriori* (MAP) estimation requires knowledge of the *a priori* distribution of the estimate but not a cost function. Maximum-likelihood (ML) estimation requires only the distribution of the observables with an implied parameter; ML estimation includes the term *likelihood*, since it reveals a best guess on the parameter from only knowledge of the distribution of the observables. An ML estimate is especially useful when limited information about the distribution exists and is typically the easiest to compute, often leading to a best estimate. Finally, Section 10.9 provides examples in order to contrast the features of important estimators.

10.2 FORMULATION OF THE GENERAL PARAMETER ESTIMATION PROBLEM

The set of observables or measurements y..... *ym* will be represented in vector notation by **y** = {y..... ym). Embedded within the observables are a set of parameters a..... *a* represented by the parameter vector a= {*a*...., a}.[2] The parameter estimation problem is to find a parameter estimate vector **&** (y) that depends on the observables and in vector representation is expressed as *ά* (ỹ) = {ά.....}. Note that the observables vector given later is suppressed in **a** for simplicity.

[2] In this chapter, the parameter vector refers to the set of parameters to be estimated; it does not refer to the received- signal amplitude.

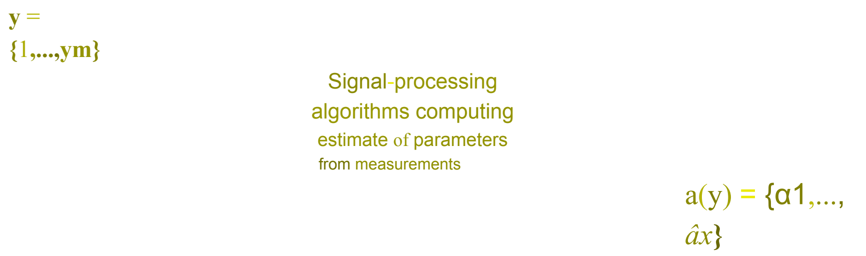


Figure 10.1. Parameter estimation from a set of measurements

Figure 10.1 pictorially describes the parameter estimation problem. Concepts that need to be understood at this point are as follows:

1. The parameter estimate (y) depends on the random observables and is therefore itself a random variable (or random vector).

2. The parameter & may be a random vector or a nonrandom parameter vector. If **a** is a random vector, the relationship between the random measurements **y** and ***a*** is described in terms of the conditional pdf $p(a)$. If a is a nonrandom parameter vector, the pdf of the measurements is written to identify the nonrandom parameter vector by $p(y;$ a). In this case a pdf for a is not defined, since a is a vector of nonrandom values. Specific cases of parameter estimation occur depending on whether the pdf of ***a*** is known or unknown. 3. This note is an extension of the prior note, particularly suitable for the mathematical calculations in this and the other estimation-theory chapters. It is often necessary to perform mathematical procedures such as differentiation, integration, and basic algebraic manipulation with respect to the parameters a. When a is a nonrandom parameter or when a pdf is conditioned upon a, these procedures are legitimate. Throughout these chapters we have used boldfaced type when the parameter refers to a random variable and normal type when it refers to a mathematical or numerical variable.

4. The pdf of the observables is not always known, but parameter estimates may still be attainable.

Now that parameter estimates are defined, it is natural to ask how good the estimates are; *i.e.*, does the estimate characterize the parameter, and is it the best possible estimate? To answer these questions it is necessary to investigate properties of the estimate as described in subsequent sections.

Example 10.1

To clarify the foregoing abstract concept. suppose that an estimate of the mean is desired from a set of m statistically independent Gaussian random variables, each with unit variance and the same constant mean µ. The observables **y** = **{y1,..., ym}** each have a pdf described by

P(Yi; µ) =

1

√2π

√ exp(-

(1⁄2µ)2)

(Yi -

2

i = 1,..., m

(10.1)

The notation *p(y;; µ*) explicitly identifies that the mean u is the unknown nonrandom parameter to be estimated, and the underlying random variable y;, often indicated as a

subscript, is suppressed for simplicity. An estimate of the mean, denoted by , is the sample mean presented in Chapter 1, *i.e.*,

#21

µ

Σyi

m

(10.2)

Since each y, is a random variable, the estimate **µ**, corresponding to ***â*** in the general formulation, is also a random variable. Once the observable data are obtained, the estimate û is a number, *i.e.*, a specific realization of the random variable.

Since each observable y; in this example is a Gaussian random variable, the estimate ***û*** formed by a linear combination of Gaussian random variables is also Gaussian, from Appendix C. The mean of **µ** can be obtained as follows:

1

E{M} = *E*

-EΣ-Σ

Σ

τ

m

The variance of i, o, is given by

με

o2 = v

1

m

Σ

=

m

ΣE{y}} =

μ

(10.3)

ııı

i=l

ΣV{yi

}

m

(10.4)

Thus, the estimate $\boldsymbol{\mu}$ has an average value that is the true mean of the observables. The variance of the estimate indicates that with a single observable, *i.e.*, m = 1, the result would be no better than the variance associated with the original random variable. As the number of observables *m* used in the estimate increases, the variance of the estimate becomes smaller. This behavior represents desirable properties of a good estimate.

Example 10.2

In a running sum the sample mean and sample variance are defined for the current sample *i* as

and

y

52

(y

Σ 2

(10.5)

(10.6)

Using the MATLAB routine zmuv.m given in www.prenhall.com/schonhoff, 10.000 samples of a zero-mean, unit-variance normal random variable were generated. The results for the sample mean and variance are plotted as a function of the sample number *i* and shown in Figure 10.2. After 10.000 samples it can be seen that there is excellent convergence between the true and estimated values. On the other hand, the figure also shows that good convergence occurs only after a few thousand samples.

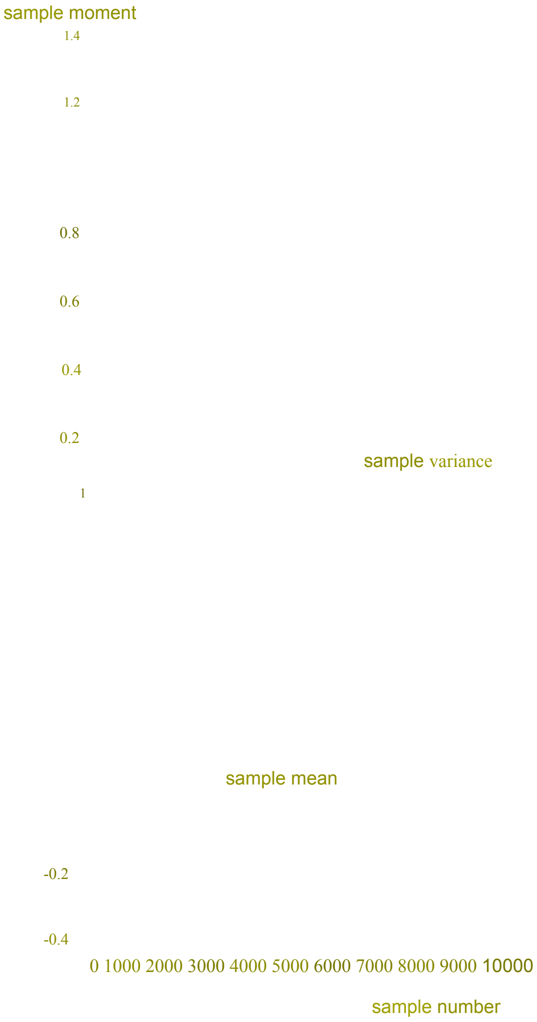


Figure 10.2. Sample mean and variance determined by MATLAB

10.3 RELATIONSHIP BETWEEN DETECTION AND ESTIMATION THEORY

Estimation and detection theory problems are similar and follow many common analysis procedures. One way of distinguishing between them can be explained as follows: In detection theory, the problem is to select from a number of hypotheses whether a signal (or situation) is present or not. In estimation theory, it is assumed that the signal is present; it is more important to obtain an estimate of a signal (or situation) parameter. Examples of parameters to be estimated might include the amplitude or time of arrival of a radio signal, or the pH of a chemical compound.

Detection and estimation theory are often accomplished together. Parameters of a signal may have discrete values, with separate hypotheses assigned to each value; thus, the presence of the signal as well as the discrete values of the signal parameters are simultaneously determined. In this case, the distinction between detection and estimation theory becomes blurred. In general, detection theory assumes a discrete set of hypotheses, whereas estimation theory assumes a continuum of values. In fact, an approach followed by many authors[3] is to treat estimation theory as a limiting case of detection theory, where the number of hypotheses becomes larger so that a continuum of possible hypotheses exists. Then, instead of making a binary or *M*-ary decision about the parameter, the multiple hypotheses are used to compute an estimate of the actual value of the

[3] **See**, e.g., [93] and [114].

[4] See Chapter 8.

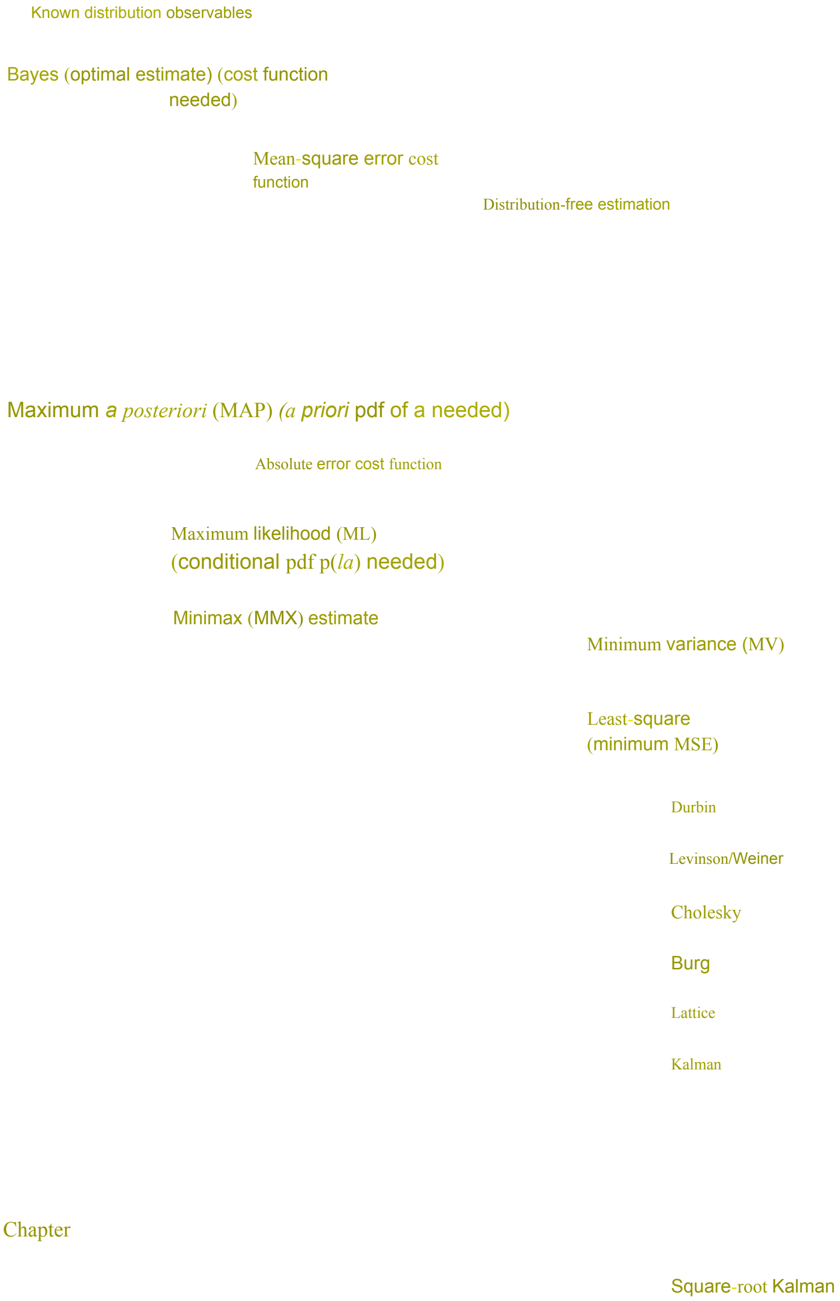
parameter from the observations. Estimation theory can be applied to provide estimates of signal parameters or, in conjunction with detection theory, used to extract values of signal parameters.

10.4 TYPES OF ESTIMATION PROBLEMS

As indicated previously, parameter estimation falls naturally into two classes corresponding to known and unknown distributions of the observables. Figure 10.3 outlines numerous estimation procedures using this classification. Many additional assumptions are required to continue detailed computation, such as

- Estimation of a constant or a random variable
- Linear or nonlinear estimates
- Continuous or discrete parameters
- Independent or correlated random variables
- Single- or multiple-parameter estimates
- Stationary or nonstationary models
- Known or unknown correlation coefficients
- Recursive or nonrecursive estimates
- Gaussian or non-Gaussian random variables

Parameter estimation


Known distribution observables
Bayes (optimal estimate) (cost function needed)
Mean-square error cost function
Distribution-free estimation
Maximum *a posteriori* (MAP) *(a priori* pdf of a needed)
Absolute error cost function
Maximum likelihood (ML) (conditional pdf p(*la*) needed)
Minimax (MMX) estimate
Minimum variance (MV)
Least-square (minimum MSE)
Durbin
Levinson/Weiner
Cholesky
Burg
Lattice
Kalman
Square-root Kalman

[5] See Chapter 2.

Figure 10.3 . Classification of parameter estimation problems

Thus, it is apparent that an exhaustive treatment of the various cases is an arduous task. This chapter focuses on estimation with a known distribution of observables, whereas later chapters treat distribution-free cases using specific assumptions extracted from the preceding list. See [45].

10.5 PROPERTIES OF ESTIMATORS

Properties of estimators are considered in this section; for simplicity, a single parameter a with an estimate **&** (y) will be postulated. Possible properties for an estimate include (see [33], [86], [93], [83], and [58]):

- Unbiased
- Consistent
- Invariant
- Sufficient
- Minimum variance
- Efficient

Asymptotically
efficient
Asymptotically normal

Each of these properties is introduced in more detail throughout the next sections.

10.5.1 Unbiased Estimates

An estimate **&** (y) for the single nonrandom parameter a is *unbiased* if *E*{*a*} = a for all a. As an example of unbiased and biased estimates, consider the case of the sample mean and sample variance formed by m identically distributed random variables yi, i = 1,..., *m*. It is further assumed that the random variables are mutually uncorrelated and each has the same mean μ and same variance o. The sample mean **y** is given by

$$\mathbf{y} \qquad \sum_{i=1}^{m} y_i$$

and the sample variance s is expressed as

$$\qquad {}^{m} \qquad (Y_i \qquad (10.7)$$

(10.8)

"Much of the discussion of properties of an estimate & is based on the specialization of a single parameter rather than a vector, so that the notation is less complicated.

The sample mean is unbiased, since $E\{y\} = \mu$. The sample variance, on the other hand, is biased as a result of the following argument:

Since

m

$E\{s\}$

}

$= E$

(yi

m

m

1

$= E$

m

(10.9)

m

0.

$E\ \{(\mathbf{y}\ ; -)(\mathbf{y}\text{-}\mu)\} = E\ (yi - \mu)$—

ΣO

i #j

(y

(10.10)

777

i = j

m

Eq. (10.9) can be expanded as

E

=

m

m

$\Sigma[\omega$ - $\mu)$ -

G

m

2 m m

$[E\{(y;$ - $\mu)2$ }

= =

$= \Sigma E$ $\{(y;$ – $\mu)(y$– $\mu)\}$ + = $[E\{(\tilde{y}$-$\mu)2\}$

(10.11)

i=1

m i=1

The first term is o3. The second term is, from Eq. (10.10),

m

m

m

2

ΣE -$\mu)$

(yi

m

m

i=1

(yj j=1

m

2

m

n2

Σo

(10.12)

m

i=1

Since the third term of Eq. (10.11) is independent of i, it can be written as

m

1 $\Sigma E\{(\tilde{y}$ - $\mu)2\}$ =

$E\{(\tilde{y}$ — $\mu)2\}$

m

m

Eyk $\mu)$

m

The combination of the terms in Eq. (10.11) results in

$E\{s\}$

$\}$

1 m

$m2$ $k=1$

1 111

$E\{(y_i$ — $\mu)^2\}$

m m

$= 0$

771

m 1

^{2}

σ

I

M

(10.13)

(10.14)

Since $E\{s3\}o$, the estimate of the sample variance is biased. An unbiased estimate for the sample variance can be obtained by including the factor *m/(m* − 1), *i.e.*,

m

E

m 1

(10.15)

In general, an estimate **&**, represents a sequence of estimates of the unknown parameter *a*. The estimate is defined to be *asymptotically unbiased* if

$$\lim_{m\infty} E\{am\} = \alpha$$

(10.16)

An example of an asymptotically unbiased estimate is the sample variance, given above, which approaches the true variance for large *m*.

An unbiased estimate is also defined in the case when the parameter **a** is itself a random vector with mean $E\{a\} = \mu a$. If the conditional mean $E\{\sim|\sim\} =$ a for all values of a, then the estimate & is said to be conditionally unbiased. Assuming *p(a)* is the known density function of the random vector **a**, the estimate ***&*** is said to be unconditionally unbiased if

where

$$E\{E\{ã|ã\}\} = E\{ã\} =$$

$$E(ã\} = \mu a$$

$$E\{E\{ã,ã\}\} = [^{\circ} E$$

$$\{ã,ã\}\, p(\bar{a})d\bar{a}$$

$-\infty$

10.5.2 Consistent Estimates

(10.17)

(10.18)

An estimate & is *consistent* if it converges stochastically to a as *m* becomes large. An explicit statement of consistency is that for every small number, € > 0, the estimate *âm* concentrates on the parameter a as *m* becomes large, *i.e.*,

$$\lim_{m\to\infty} P[a - 6\ \hat{a}m < \alpha +] \to 1$$ (10.19)

As an example, assume the random variable **y** has a normal distribution with unknown mean μ and a known variance o2. The set of random variables **y**..... ***ym***, represents observed random samples obtained from the same population. The sample mean, denoted by **y**, *i.e.*,

m

Im

Y_i

m

(10.20)

approaches the population mean u as *m* becomes large. As a result, **y**,,, is a consistent estimate of μ.

[7] Note that the subscript *m* has been added to *a* here to explicitly indicate the dependence on the number of random variables. Ordinarily this subscript and the dependence of the estimate on the observations y are suppressed for convenience in notation.

10.5.3 Invariant Estimates

A desirable property of an estimate is that it is *invariant* under a transformation of a parameter. Specifically, if **a** is an estimate of a and g(a) is a function of a, then ***a*** is an invariant estimate if g(a) is an estimate of g(*a*). Consider the sample mean $\hat{u}$ for mutually uncorrelated random variables y;, each having mean and unit variance. The square of the sample mean 2 is a function g(μ) = 2. It might be expected that, since is an estimate of . then 2 is a good estimate of u2, but that is not the case. The sample mean is unbiased, i.e., E{} = μ, but *E*{μ2} = μ2 +, so that invariance is not attained. A scale change, such as pounds to kilograms or seconds to minutes, is an example where invariance is satisfied. Let us reexamine the sample mean ***μ*** of the random variables y;, i = 1,..., m discussed previously. If **a** is a scalar, then **au** is an estimate of au. As indicated in Section 10.8, maximum-likelihood estimates do have the property of invariance.

10.5.4 Estimators Based on Sufficient Statistics

Let *T*(), termed a statistic, be a function of the observed random variables y1...., *ym* that are generated by a pdf p(1, 2, ... ym: a) with a parameter a that is not directly observable. When random samples are obtained that are dependent on the parameter a, the statistic computed from the random samples is used to make inferences about this parameter. It is then desirable to use a statistic that summarizes all of the knowledge available in the measurements in a more manageable form (i.e., a single random variable rather than m random variables) and therefore can be a good estimator of the parameter a. For example, an estimate **a** = y does not contain all of the information about **y**, but the sample mean ***a*** = *T*(y) = y, in the case of independent. identically distributed normal random variables, does retain all of the information and is therefore a sufficient estimate. This leads to the definition of a sufficient statistic. *T*() is a sufficient statistic for a if *p*(y, y2,... *Ym* T()), the conditional pdf of the random variables y. y..... 'm given the value of *T*(), does not depend on a from [33] and [86].

A method used to determine whether or not an estimate (or statistic) is a sufficient esti- mate is to see if the Fisher factorization theorem is satisfied. The theorem states that if the pdf *p*(y1 y2,..., *Ym*; a) can be factored into a function g(*T*(), *a*) that depends on the measurements through the statistic *T* (y) and the parameter a and another function *h*(yi, y2,..., ym) that depends only on the observables, then *T*() is a sufficient statistic. Conversely, if *T*() is a sufficient esti- mate for the parameter a, then the pdf *p*(yi, y2.... : *a*) can be factored. Mathematically, the Fisher factorization theorem is written as

$$p(y_1 \ldots y_m; \alpha) = g(T(y1, \ldots, ym), a)\, h(y1, \ldots, Ym) \tag{10.21}$$

where g() is a function that depends only on the estimate and the parameter a and *h*() is a function of the observables and is independent of the parameter *a*.

An intuitively satisfying way of explaining sufficiency is as follows: We know that a affects the observations y1 y2...., Y; thus, we learn about a by viewing y and observing its statistical behavior. If knowing g(T(), a) removes any further dependence on a of the distribution of **y**,

[8] See Problem 10.4 for proof.

it is clear that g(T(), a) contains all the information in p(yi, y2.... Ym: a) that is useful for estimating a. This condition is why it is called *sufficient*. Since the conditional distribution of y given T() does not depend on a, then neither does the conditional expectation of a function of y, denoted as $D(\tilde{y}) = E\{\tilde{y}|T\ (\mathring{y})\}$. Furthermore, this conditional expectation is typically a better estimator of a. As a result, an important point in regard to sufficient statistics is that if two unbiased estimates are available, one based on a sufficient statistic and one not, the estimate based on the sufficient statistic will have a smaller variance. This estimator is known as the Rao- Blackwell estimator from [167]. Thus, a minimum-variance unbiased estimate of a parameter is based on a sufficient statistic. See [86] and [33].

As an example of a good estimator based on a sufficient estimate, consider the sample mean **y**, where the individual random variables **y** are independent and each is normally distributed with mean μ and variance o2. The joint pdf of **y1**,..., ***Ym*** is given by

$$p(y1, ..., ym; \mu, o2) = \frac{1}{(2\pi\sigma2)m/2} \exp(-\sum_{i=1}^{m} \frac{(Yi - \mu)2}{202}$$

The preceding pdf can be expressed in a different form by using the identity

$$\sum_{i=1}^{m}(vì - \mu)2 = \sum_{i=1}^{m}[(y; - y) + (\tilde{y} - \mu)]2$$

$$= \sum_{i=1}^{m}(yi - y)2 + m(\tilde{y} - \mu)2 + 2\sum_{i=1}^{m}(y; - \tilde{y})(\tilde{y} - \mu)$$

where

and

$i=1$

m m

Yi

$(yi$

$\Sigma - D$

m

$i=1$

$i=1$

(10.22)

(10.23)

(10.24)

(10.25)

Since the last term in Eq. (10.23) is zero, the pdf in Eq. (10.22) can be rewritten, using Eq. (10.8), as

$p(y1, . . ., Ym; \mu, o2) =$

exp

$-m$

202

exp -m

202

y

1

(2л02)m

/2

(10.26)

Notice that the first factor involves only the parameter μ and its estimate **y**; the second factor involves only the estimate s. In this case the pdf of y depends on the observations only through the value of y and s3, where *h(y,...,* *Ym)* = 1. Therefore, the sample mean y is a sufficient statistic for μ, the mean of the distribution and the sample variance s is a sufficient statistic for the variance o2.

10.5.5 Minimum-Variance Estimates

In estimation it is desirable to seek an estimate & that is close in some sense to the parameter & being estimated. This behavior is true even if the estimator is biased. For an unbiased estimate the variance, expressed as *V{a}* = *E*((&- a)}, is a measure of variability of an estimator about the parameter a and is, in fact, the mean-squared error between the estimate and the true value of the parameter. If &1 and 2 are two unbiased estimates of the parameter & and *V*{2} ≤ *V*{â¦}, then it is natural to use an estimate that attains the smaller variance, *i.e.,* V{2}.

It should be noted that unbiasedness is a good property of an estimator, since a biased estimator may lead to an unrealizable estimate. However, biased estimates do exist that have smaller mean-squared error than unbiased estimates, although these estimates are typically asymptotically unbiased. See, for example. [69]. Figure 10.4 shows a depiction of the pdfs of two unbiased estimates for a labeled &, and 2. Clearly 2 has a smaller variance than & and is therefore a better estimate.

An important bound on the variance of the estimate is the Cramér-Rao inequality for an unbiased estimate; it results in a minimum-variance estimate when the equality in the bound is attained. To present the Cramér-Rao inequality, let **y**.....**y** be independent random variables with the same pdf p(y;; a) and assume that an unbiased estimate **&** exists so that E {a} = α. Then one form of the Cramér-Rao inequality is stated as

V {*a*} >

1

a2 *E* ln *p*(yi

da-

Ym:

(10.27)

The denominator of the last equation is known as the Fisher information *F(a)* from [69]. Inequality (10.27) can then be expressed as

where

$$V\{a\} > F^{-1}(\alpha)$$

22

$$F(\alpha) = -E \quad da2 \quad \text{In } p(\text{y} \ldots ym\ \alpha) \qquad (10.28)$$

(10.29)

In this form it is seen that a larger value of the Fisher information F(a), where more information is known about the estimate, implies a smaller variance. If the estimate of the parameter satisfies the Cramér-Rao bound with equality (i.e., it is a minimum-variance unbiased estimate), then it is

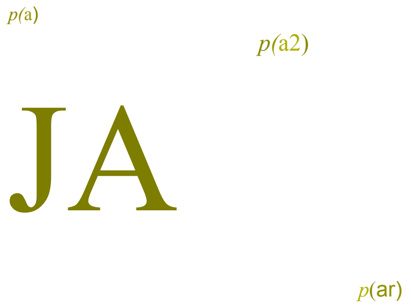


Figure 10.4. The pdf of two unbiased estimates

termed a *most efficient estimate*. It should be noted, from [164], that an efficient estimate is also a sufficient statistic.9

An alternative form of the Cramér-Rao inequality that is convenient for use in determining the minimum variance of an unbiased estimate is given by

$$V(\alpha) \geq \frac{1}{E \quad \text{a In plyi} \atop \theta\alpha}$$

The equality condition in Eqs. (10.27) and (10.30) is attained when

(10.30)

$$\frac{a \ln p(\text{y}1, \ldots, \text{ym}; \alpha)}{\theta\alpha} = F(\alpha)\,[\alpha\text{-}\alpha]$$

10

(10.31)

11

where $E\{a\}$ = & ( )p(a)dy and10 F(a) is a constant that is independent of y and â.11 A more general form of the Cramér-Rao inequality exists for a function of the parameter a. The Cramér-Rao inequality then takes the form, from [164],

V{2} = | [C) –

v) p; a)dy 2

dy (a)

da

a In py;α)

$\theta\alpha$

(10.32)

Equality is attained when

a

θα

In $p(y;\, a) = F(\mathrm{a})\, [\hat{a} - y(\mathrm{a})]$

A simple proof of this form of the bound is as follows:

By definition,

√ p(ỹ;

a) dy =

1

Taking the partial derivative of both sides, we get

(10.33)

a

$\theta\alpha$

=

[PG; adỹ

$= \sqrt{}$ 3PG;

α)

[

(y; a)

$\theta\alpha$

$dy = 0$ (10.34)

But the integrand is

$dp(\tilde{y}; \alpha)$

θα

$\theta\alpha$

$p(y;$ − [a

in p(5:

a)].

5; α)] ·

p(ÿ; $\alpha)$

(10.35)

[9] However, a sufficient estimate is not necessarily efficient.

[10] To simplify the notation, it is often convenient to express $p(y1,..., ym;$ a) as p($\tilde{y}$; a). In later sections, **a** is considered a random variable so that the conditional pdf $p(a)$ is used. When $\hat{a}$ appears in an integral with respect to y, the explicit dependence â = â (y) is shown.

[11] See [164].

so we can write

$ap(y;$ α)

$\theta\alpha$

Multiplying both sides by $($a), we get

In

dj

= [

ə In *p*(y; a)

θα

P(; a)dy = E

ə In *p*(y; *α*)

θα

0

(10.36)

(a) *E*

{alm p(7; α)   } =

E (v (a) In p(); α).

ə a)

=0

(10.37)

θα

Since (a) = *E*{} = *[* & *(j)p(ỹ; a)dy*,

ψ(α)

θα

= √

a(5)

dp(ỹ; α)

θα

-

a *(y*

ə In *p(y;* α)

θα

p(y; a)*dy*

(10.38)

Subtracting (10.37) from (10.38), we get

[[a (5) - 1

(a)] a la p(F;

α) [[â(5) − ¥

θα

dy (a)

p(ỹ; a)dẏ =

(10.39)

da

The Schwartz inequality for functions *f*(x) and *g*(x) can be written as

|f

f(x)g

(x)

f(x)*g*(*x*)dx <

≤ [\f (x) P2

dx f 18(x)

12 *dx*

where equality holds if f(x) = *kg** (x), with *k* being a constant. If we square Eq. (10.39), we get

2

dv ((@))] = [[[ã(3) -

v(**a**)] **31n** p6: a) p();

ady]"

(dy

da

In *p(y;*

θα

Using the definition of the variance from Eq. (10.32) and choosing

f (ÿ) = [â(ÿ) − ↓ (*α*)] √p*(ỹ;*

a)

(10.40)

(10.41)

and

8() =

ә In p(y; α)

θα

√p(y; α)

(10.42)

in the Schwartz inequality, we get the desired result shown as the inequality of Eq. (10.32).

Now we want to return to the form of the Cramér-Rao bound given by Eq. (10.27). To prove this form, take the partial derivative of Eq. (10.36) with respect to a, *i.e.,*

√

[3

2

a2 In *p(y; α)*

да2

p(ỹ; α) +

θα

a ln *py*; a) *ap(y*; a)

θα

dy=0

From Eq. (10.35) and the expectation of the preceding equation, we obtain the following:

E

ә In *p(y*; a)

θα

-E

да2

{302 In

p; a)}

(10.43)

(10.44)

In the unbiased case where (a) = a, Eq. (10.32) reduces to Eq. (10.27). The Schwartz inequality condition can now be stated as

$$\hat{a} - \surd(\alpha) = K(\alpha)\text{-}\frac{\partial}{\theta\alpha}\text{In } p(y;\alpha) \qquad (10.45)$$

where (α), in general, depends on a. Computing the partial derivative of the previous equation with respect to a results in

$$\frac{ay(\alpha)}{\theta\alpha} = K(\alpha)\text{-}\frac{a2 \text{ In } p(y;a)}{да2} + \frac{JK(\alpha)}{\theta\alpha}\text{ In } p(y;a) \qquad (10.46)$$

Forming the expectation of this equation and using $E\{\frac{\text{In } p(y;a)}{\theta\alpha}\ ak(\alpha) = 0)$ from Eq. (10.36) results in

$$K(\alpha) = \frac{\frac{\lambda\psi(\alpha)}{\theta\alpha}}{\frac{22 \text{ In } p(\tilde{y};\alpha)}{да?}} \qquad (10.47)$$

In the unbiased case where (a) = α, κ (α) is seen to be the reciprocal of the Fisher information $F(a)$. Replacing $\kappa(a)$ by $F\text{-}1(a)$ in Eq. (10.45) proves Eq. (10.31).

A special case of Eq. (10.32) exists for biased estimates. If b(a) represents the bias of the estimate where $E\{a\}$ = a + b(a), then an alternative form of the Cramér-Rao bound for biased estimates is obtained using Eqs. (10.44) and (10.32) and can be written as

$$V\{\mathbf{a}\} > \frac{\left(1+\frac{db(a)}{da}\right)^2}{E\left[\frac{22 \text{ In } p(y;\alpha)}{да2}\right]}$$

When the bias b(a) is zero, the foregoing equation reduces to Eq. (10.27).

(10.48)

Another result applicable to a linear function of a is that efficiency is maintained in the transformation. Suppose the linear transformation is estimate of a, then (a) is an efficient estimate of expectation of *(a)*, i.e.,

(a) = = aa + b. Thus, if & is an efficient (*a*).
To prove this result, compute the

$$E\ (\mathrm{v(a)}\} = E\ \{a\acute{\alpha} + b\} = aa + b = \surd(\alpha)$$

(10.49)

The variance of the estimate of a linear function is given by V {(*a*)} = *V* {***aά*** +b} = *a2V* {a} . From the Cramér-Rao bound the minimum variance is attained, since

$$V\ \{\mathrm{y}\ (\boldsymbol{\alpha})\} >$$

$$\mathrm{dy}\ (\alpha)$$

$$\theta\alpha$$

*

$$F7\ (\mathrm{a}) = a2V$$

{â}

(10.50)

Note that nonlinear transformations do not, in general, exhibit this property, but do approach this result asymptotically for long data records.

Example
10.3

An example of a most efficient estimate is obtained by once again considering the independent, normally distributed, random variables **y1**, ..., ym with mean μ and variance 2. The sample mean y is an efficient estimate of the mean μ, as shown in the following argument. The sample mean is unbiased and its variance is V{y} = ○/m. The right-hand side of the Cramer-Rao inequality is evaluated using the identity provided in Eq. (10.23), *i.e.*,

$$\mathrm{In}\ p(\mathrm{y}.....\ \mathrm{Ym};\ \mu) = \mathrm{In}\ \frac{1}{(2\pi\sigma2)\mathrm{m}/2}\ \mathrm{exp}$$

202

771

i=1

(yi

[(Y -

1)2)]

771

1

In(2лo2

)

2

202

Σou –

μι

i=1

ıı

77l

2

Σ

In (270)-(-5)2-1-) (10.51)

i=1

The right-hand side of the Cramer-Rao inequality is *F*) and can then be computed

as

F-1(u) =

"

82 дн

1

-1

E|(-2))

m

-*E* **In** *p*(y1、 ... ...,

Ym; ft) } E

(10.52)

Therefore, the inequality is satisfied with equality in this case and the sample mean **y** is a most efficient estimate for u, the mean of the distribution.

Example 10.4

As an example of an estimate that is not efficient, consider the sample variance described by Eq. (10.8), *i.e.*,

Jı

i=1

(10.53)

where the random variables y..... ym are each independent and normally distributed

ms with mean μ and variance o2. Note that is chi-squared with (m-1) degrees of

freedom. From Table B.1 in Appendix B, the mean and variance of are given by

ms

and

E

= m - **1**

σ

(10.54)

ms

= 2(m − 1)

(10.55)

Therefore, the mean of s is

E(s) = (m =

1) ∘ ('

σ'

2

m

(10.56)

so that s is biased. The variance of s is given by

V

{s3

}}

2*(m*-1)

4

Let a function *f* be defined as

m2

σ

(10.57)

f = ln *p*(y1, ..., Ym;
o2)

m *m*

1

-

In(2πσ2

)

2

202

Σω - μ)2

(10.58)

i=1

For simplicity, replace o2 by s so
that

Then, from Problem 10.6,

a2 f

∂2

m

252

1 m

Σ(yi -
μ)2

(10.59)

i=1

m

E

∂2

204

(10.60)

and the right-hand side of the Cramér-Rao inequality becomes

204

9 m

of s does not equal the estimate is not efficient.

204

Since the variance

771

Since this example is one of a biased estimator, the more general bound of Eq. (10.48) should be used. The numerator in (10.48) can be evaluated from the sample variance s using

$E\{s\}\} =$
o2

m

(10.61)

Therefore, the bias is -o2/m and the right-hand side of the inequality (10.48) can be written as

(1 -
4)2

m

m
204

m
204 (m1)2

$m2$

m -

m

(10.62)

Thus, the bound on the variance of the biased estimate on the right-hand side of the Cramér-Rao inequality is smaller than the variance of s by the factor $(m-1)/m$.

Example 10.5

Now, consider the case of a known mean with an estimate of the variance given by

mn
1
$0 =
Σ(vi -
μ)2
m
$i=1$

(10.63)

ms
o

Note that is chi-squared with m degrees of freedom and a mean and variance given by

and

mso

E

= *m*

V

mso

02

= *2m*

(10.64)

(10.65)

The estimate s3 is unbiased, since *E*{s}} = σ2, and it is also most efficient, since *V*{s}}

204

m

corresponding to the equality condition in the Cramér-Rao bound.

Example 10.6

A final case in this section to be considered is the estimate of the variance given by

In this case

(m-**1**)**s?**

62

are given by

772

1

m

1

Σ(yi -

y)2

i=1

(10.66)

is chi-squared with *m* - 1 degrees of freedom; the mean and variance

E { (m

= 1s] }

$= m$ I

02

(10.67)

and

V

σ2

v { (m−Ds] } =

2(n-1)

=2(*m*

(10.68)

As a result

and

$E\{\mathbf{s}\} =$

o2

204

$V\{s\}$

}

m

–

1

(10.69)

(10.70)

This estimate is unbiased but not efficient, since the Cramér-Rao bound is not attained.

10.5.6 Asymptotically Efficient Estimate

Let ao and *â* be unbiased estimates of the parameter *a*, and assume that do is the most efficient estimate. Then a measure of efficiency, &, can be defined as the ratio of the variances of the estimates, *i.e.*,

ε=

V {**ão**} *V{**α1***}

<1

(10.71)

In the preceding example for estimating the mean using y, it was determined that the Cramér-Rao bound was o2/m and, since the variance of y is also o2*/m*, ε: 1. For the cases involving an unbiased estimate of the variance, it was determined that the Cramér-Rao bound was 204*/m*. For so the variance is 204*/m* and for s the variance is 204*/(m* - 1), so that

ΕΞ

An estimate is termed *asymptotically efficient* if

m 1

m

lim & = 1

Therefore, the estimate s is asymptotically efficient.

10.5.7 Asymptotically Normal Estimate

(10.72)

(10.73)

An *asymptotically normal* estimate is one that yields a normal distribution of the random variable √*m(*α-a) as m approaches infinity. A best asymptotically normal (BAN) estimate occurs when the √*m(*â*m* − a) is normal with zero mean and a variance that is smaller than any other asymptotically normal estimate.

m

As an example of a BAN estimate, consider samples drawn from a normal distribution with mean μ and variance o2. Then the sample mean

172
1
y_T
Yi
m
(10.74)

is an efficient estimate and a BAN estimate. The distribution of $\sqrt{m}(\mathbf{y})$ is normal with zero mean and variance o2, and no other estimate can have a smaller variance.

10.5.8 Extensions of the Cramér-Rao Bound

The first extension of the Cramér-Rao bound described here considers the case where the parameter *a* is a set of *L* parameters represented in vector form as a = (α, ..., α). An unbiased estimate ***a*** = (a1, ..., *άL*) is then developed where each parameter a; has a corresponding estimate *â¡*. The minimum-variance unbiased estimate in this case is then described in terms of an L × *L* matrix referred to as the *Fisher information matrix*. The bound is described more completely

in Section 11.9, where simultaneous parameter estimation is developed. An important general conclusion is that the Cramér-Rao bound increases as more parameters are estimated.12

Another generalization of the Cramér-Rao bound considers the case where the parameter ***a*** is a random variable with pdf *p*(a) and the estimate & (y) is unbiased, *i.c.*, E{â|} = *E*{a}. In this case, the Cramér-Rao inequality given by (10.30) becomes a bound on the minimum mean-square error expressed in terms of the joint pdf *p*(y1, ..., ym, α) as

1

An alternative form is

E €
{[â(3) −
α]2 }
>
E
© {[α (5)
− α]2 }
Σ
(10.75)

θα
E{[In *p*(y1, *ym*, *α*)

$$-1$$

(10.76)

$$\text{да2}$$

$$E \{322 \text{ In } p(y1, \ldots, ym, \alpha) \quad \alpha \quad)\}$$

with equality if

$$\text{In } p(y1, \ldots, ym, \overset{a}{\alpha}) = F(\hat{a}(\tilde{y}) - \alpha) \ \theta\alpha$$

(10.77)

In the foregoing equation, the Fisher information *F* does not depend on the random variable *a*. which has been averaged out in the bound computation. References [152] and [90] provide the derivation of the unconditional bound of inequality (10.75). A further extension of this bound described in [90] is available for the case where a is a random vector.

10.6 BAYES ESTIMATION

Classical estimation theory involves the estimate of unknown but deterministic parameters. How- ever, it is often possible to model the parameter to be estimated as a random variable a with some postulated distribution p(*a*), referred to as the a *priori* pdf.13 For example, suppose there are many values for the sample mean collected over some period of time. Bayesian methods allow introduction of this data *via* the assumed distribution of the sample mean. There may also be a weighting function that incorporates the cost *C*(a, *ά*()), also referred to by some authors as a loss, introduced by the error *a* between the estimate and its true value and defined by

$$de = \boldsymbol{\alpha} \rightarrow \hat{\mathbf{a}} \ (\tilde{y})$$

(10.78)

An example of such a weighting function is the mean-squared error between the estimate &() and the parameter a. This cost is the counterpart of that observed in detection theory as described

[12] See, e.g., Section 11.9 or reference [69].

[13] For simplicity, a single random parameter is assumed; in general, there may be a set of random parameters to be estimated.

in Section 4.3. Similarly, an average risk R is obtained by averaging the cost over p(a.), the joint pdf of the parameter **a**, and the observable **y** according to

R = E{C (α, & (y))) = [' ~

[~ C (a, & (j))p(a, j) da đỷ

8

Minimizing this average risk yields a Bayes estimate.

14

(10.79)

An alternative expression for the average risk can be obtained in terms of the conditional cost Co(**a**)ly), defined by

8

C(@\ỹ) = √ ° C.(*α,*

â (5)p(a\ỹ) *da*

1

8

so that the average risk can be expressed as

(10.80)

∞

R

[

p)C(@(5

)l) dỹ

(10.81)

Since the pdf of the observable is positive over the range of its outcomes, minimizing the average risk is accomplished by minimizing the conditional risk.

A different and interesting way of introducing Bayes estimation is to recognize that mathe- matically, the estimate is determined from Eq. (10.80) and can be written as

∞

άB = min E {C (α, **â**)} =

C(a, a)*p*(aly) *da*

81

The *a posteriori* distribution is obtained from Bayes' theorem and can be written as

$$p(aly) =$$

$$p(a)p(a)\ \mathrm{L}p(a)p(a)$$

$$da$$

(10.82)

(10.83)

where $p(a)$ is referred to as the likelihood distribution, since it provides the current observations reflecting the value of the parameter a.

From the previous paragraph, it can now be seen that a Bayes estimate is based on *a posteriori* distributions where it is necessary to know the *a priori* distribution and the likelihood distribution. The *a priori* distribution represents the knowledge of the parameter before any data are available. Following observation of some data, the current knowledge may change, resulting in a change to the *a posteriori* distribution. Thus, Bayes estimation can be viewed as a sequential *learning algorithm*. If a second independent observed data set appears after y is measured, then a sequential version of Bayes' theorem can be written as

$$p(ax, y) =$$

$$\infty$$

$$p(x,ya)p(a)$$

$$p(x, ya)p(a)\ da$$

$$8$$

$$p(xa)p(a)p(a)$$

$$p(x,ya)p(a)da$$

$$p(a)\mathrm{p}(a)$$

$$p(ax)p(\tilde{y}|a)\ da$$

(10.84)

[14] A more conservative strategy would be to minimize the maximum risk, resulting in minimax estimation.

This last expression indicates that the a *posteriori* distribution of a after observing the new information can be interpreted as an extension of Eq. (10.83), where is now the *a priori* information.

Bayes estimation assumes that the probability distributions are known and that the integrations can be computed. Thus, simple a *priori* and likelihood distributions referred to as *conjugate pairs* are sought. For example, in estimating the mean, a Gaussian distribution is used for both the *a priori* and a *posteriori* distributions.

To proceed further, a cost function must be selected. Examples of cost functions include:

1. Squared-error cost function, defined by

$$C_s\ (\alpha,\ \hat{a}\ (\tilde{y})) = (\alpha - \hat{a}\ (\tilde{y}))2 = a2 \qquad (10.85)$$

2. Uniform cost function defined by

$$C_u\ (\alpha,\ \hat{a}\ (y)) = \left\{ \begin{array}{ll} 0, & J\alpha el < A \\ 1, & 1\alpha el > 2/2/2 \end{array} \right. \qquad (10.86)$$

where A is a small number

3. Absolute-error cost function, defined by

$$C_{AE}\ (\alpha,\ \hat{a}\ (y)) = |\alpha el \qquad (10.87)$$

These cost functions are depicted in Figure 10.5 as a function of the error ae. See [122], [93], [152], [57], and [129].

10.6.1 MSE Minimization

Using the squared-error cost function in the conditional cost equation and averaging over the conditional pdf of **a** allows the conditional cost C_{Ms} (aly) to be written as

$$C_{MS}\ (\alpha\ y) = \int^{\infty} (a - \hat{a}(\tilde{y}))2p(al\tilde{y})da \qquad (10.88)$$

$C_s(\alpha e)$

$C_u(\alpha e)$

$C_{AE}(\alpha e)$

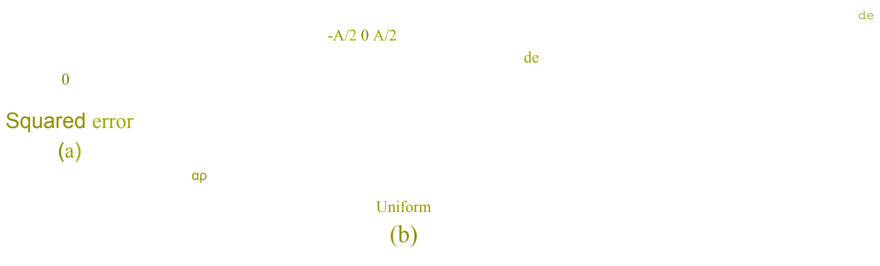


Figure 10.5. Sample cost functions

Differentiating the previous equation with respect to *â* and setting the result equal to zero yields'[15]

$$0 = \int_{-\infty}^{\infty} 2(x - \hat{a})p(a|\ddot{y})da$$

Separating the terms in the preceding equation leads to

$$\int_{-\infty}^{\infty} \hat{a}p(a|\ddot{y})da = \int_{-\infty}^{\infty} ap(a|\ddot{y})da$$

Since fp*(aly)da* = 1 and â does not depend on a, this equation can be written as

$$\hat{a}_{MS} = \int_{-\infty}^{\infty} ap(a\backslash 5)\, da = E(\alpha\{\}) \qquad (10.89)$$

(10.90)

(10.91)

Note that the subscript *MS* is added to the parameter â to indicate that a mean-square error estimate (MSE) has been obtained. Note that the result is a minimum, since the second derivative with respect to â is positive. Equation (10.91) is the conditional mean of the parameter a, given the observations **y**. It depends on the *a posteriori* pdf *p(*ay) and is independent of the pdf of **ỹ**.

An alternative form of Eq. (10.91) can be obtained by using the identities

and

p(aly) =

P(ỳα)*p(a)*

P(y)

P(3) = [

p (3\a)p(a)da

p

resulting in

(10.92)

(10.93)

fap*(a)p(a)da*

â

MS

fp(*a)p(a)da*

(10.94)

This form requires the pdfs *p(*ya) and *p*(*a*) instead of the *a posteriori* pdf *p*(*alỹ*) to compute

Ms and is often simpler to evaluate.

As shown in [93] and [164] for the following two cases, the conditional mean is an optimal estimate for a more general class of cost functions. The first case of an optimal estimate occurs when the cost function is convex and symmetric in the error de, *i.e.,* $C(\alpha, \hat{a}) = C(\alpha)$, where $C(\alpha e) = C(-\alpha e)$. A second case occurs when the cost function $C(a)$ is symmetric and nonde- creasing, *i.e.*, $C(\alpha e) > C(\alpha ej)$, for $de \geq de \geq 0$, with the *a posteriori* pdf $p(aly)$ symmetric about its mean, unimodal, and having

$$\lim C(a)p(\alpha ely) = 0 \qquad (10.95)$$

8

7

0

0

[15] Once again the notation is simplified by dropping the dependence of $\hat{a}$ on **y**.

Example 10.7

To illustrate the concept of Bayes estimation with an MSE cost function, assume that the mean μ is to be estimated by $\boldsymbol{\mu}$, using the set of observations **y1,..., ym**. Assume further that the observations are statistically independent, normal random variables, each with unknown mean μ and known variance o2. The conditional pdf pμ) is then

$$p(y|\mu) = \frac{1}{(2\pi\sigma 2m/2} \exp \frac{1}{202} \sum_{k=1}^{m} o\kappa - \mu)? \qquad (10.96)$$

Next, assume that the *a priori* pdf $p(u)$ is normal with mean m1 and variance *B2, i.e.*,

1

P(μ) =

√27g: exp (_

(id = m) 2)

2πβ

262

(10.97)

Using the identities of Eqs. (10.92), (10.93), and Exercise 4.2, the *a posteriori* pdf can be written as

p(μly)

1

√2ays exp

(_ (n =

720)2)

2πγ2

222

(10.98)

where

and

3

=

my

1

+

+

(10.99)

B

2

(10.100)

y_K

m

k=1

The Bayes estimate can now be obtained from Eq. (10.91) as

$\mu MS =$

-

м

1

exp

$2\pi\gamma 2$

= y2w =

ß2ỹ

+m1o2/m

B2 + o2/m

It follows that the random estimate Ms can be written as

μM

S

B2y+$m102/m$

B2 + o2/m

dμ

(10.101)

(10.102)

(10.103)

Computing the expected value of this estimate, assuming that the parameter μ is held constant, results 16 in

$E\{\mu MS$\μ$\}$

B2E{\µ}+m102/m

B2 +02/m

(10.104)

But the expected value of the sample mean **y** for constant µ is *E*{y} = µ, so that

E{µMS\µ} =

B2 +02/*m*

(10.105)

Since the expected value *E{*Ms} is not equal to µ, the estimate is biased. However, for a large number of observations-*i.e.*, as *m* becomes large--the estimate is asymp- totically unbiased. To see this, we take the expectation of Eq. (10.105) over the random variable *µ*, which can be stated as

E{E{µMS*µ*

}}

B2E{u}+m1o2/m

B2 + o2/*m*

where the first expectation is with respect to *µ*.

= m1 = *E*{u}

(10.106)

This last result is computed for a special case but is actually true in general. To prove this result, the expected value of the estimate is obtained, *i.e.,*

E{α*MS*} =

8

E, {*E*\&\y}} = [~ [~

ap(a\ÿ)p(ỹ)dÿda

∞

= [

ada [

[_~_ p(a, ÿ)d} =

[__ ap(a)da = E{a}

81

(10.107)

From Eq. (10.78) and the foregoing result, it can be seen that E{ae} O for the MSE cost function. Therefore, the conditional cost in Eq. (10.80) is the conditional error variance, and the average risk in Eq. (10.81) is the variance of the estimation error or error variance. The relationship can be more easily seen by writing the variance of the estimation error V{a} as

V{αe} = E {[αe - E{α,}]2} =

E{a} }

∞

= [~ [°^ [a – â

(5)12 p(x, ÿ)dadÿ

–

(10.108)

This equation is actually the average risk with an MSE cost function. Since the error variance is minimized using the MSE cost function, the estimate obtained is denoted &Ms and is referred to as a minimum error-variance estimate.

[16] In subsequent sections where there is a chance of confusion, the conditional dependence on the parameter is explicitly indicated.

Example 10.8

Example 10.7 can be continued by substituting the estimate into Eq. (10.79) to obtain the average risk, Rmin, *i.e.*,

Rmi
n

=

[_ [_μ-A)2p{u}\\p

{3}dμd}\

٢٢٠

٠

∞

8

8

p(y)

Γ

8∞

(y)du

dy

1

(μ — y2

w)*2.*

exp

2πγ2

[μ - y2

w]2 272

dudy

= [_____

p(5)y2 dÿ =

y2

-∞

B2

1

m

B2 +02*/m*

(10.109)

This expression is the minimum variance of the estimation error. An alternate derivation of the preceding equation can be obtained by computing the variance of the estimation error using

$$V\{\mu c\} = E\{\mu\}\} -$$
$$[E\{\mu e\}]2$$

(10.110)

where μμ- Ms. Note that

$$E\{\mu 2\} = E\{\mu - \mu ms\} = E\{\mu\} — E$$
$$\{E\{\hat{u}ms\backslash\mu\}\}$$

= mm 0

(10.111)

The last line of this equation follows from Eq. (10.106).

The error variance is then

$$V\{\mu,\} = E\{\mu?\} = E\ \{E\{(\mu —$$
$$\mu ms)2\backslash\mu\}\}$$

(10.112)

Examining the inner expectation results in

$$V\{\mu 2\backslash\mu\} = E\{(\mu — \mu Ms)2\ |\mu\} = E\ \{(\mu 2 —$$
$$2\mu\mu MS + \mu s)\ |\mu\}$$
$$= \mu 2 - 2\mu E\ \{\mu MS\backslash\mu\} +$$
$$E\{\mu 2\}us\backslash\mu\}$$

(10.113)

Since

$$PM$$
$$S$$

$$84y2+2m,\ B2\ (o2/m)\ y + m2$$
$$o4/m2\ (B2 + o2/m)2$$

(10.114)

it follows that

$$E\{\mu s\mu\}=$$

$$B4E\{2\backslash\mu\}+2m1B2\ (o2/m)\ E\{\backslash\mu\}$$
$$+mo+/m2$$

(10.115)

$$(82+o2/m)$$

2

Recognizing V{\μ} = o2/*m*, where *E*{y} = *μ*, it can be seen that *E*{2} 2+μ2, so that

the previous equation can be written as

B4

m

(B2μ+*m102*/m) 84 (222 +μ2) + *2m1ß2*(o2/m)*μ*+*m}oa*/m2

V{μ2\μ} = Eμ2 - 2μ.

ε

{u

2

B2+02/*m*

+

m

(B2 + o2/m)2

μ

(10.116)

and, eliminating the condition on μ, the equation becomes

V{μ} = Eμ – 2μ

(B2 μ+m102/m), B4 (1/2 +μ2) + *2m*, *B2* (o2/*m)μ* + m2œa/*m2*

+

B2 +02/*m*

m

(82 + o2/m)2

Using $E\{\mu 2\} = B2 + m2$ and $E\{\mu\} = m1$, the preceding equation can be reduced to17

$V\{Me\} =$

$B202/m$

$B2 + o2/m$

(10.117)

(10.118)

This result is the minimum variance of the estimation error expressed as Rmin in Eq. (10.109). Note that the variance of the estimate *VMs*) can also be computed, *i.e.*,

$V\{\hat{}MS\} = V$

p2y+m102/m

$B2 + 02/m$

$B2$

02

B2 +

o2/m

$V\{\mathbf{y}\}$

$B4$

$m\,(B2 +$

$o2/m)2$

10.6.2 Maximum *a Posteriori* (MAP) Estimation

(10.119)

In this section, we see that MAP estimation is a special case of Bayes estimation, where the cost function is uniform or, equivalently, unavailable and assumed uniform. The implication is that an *a priori* distribution for the parameter a is known. A MAP estimate is obtained conceptually by maximizing the pdf *p(*aly) with respect to the parameter a. A formal description is given by referring to the relationship

$p(aly) =$

$p(a)p(a)$

$p(y)$

(10.120)

The MAP estimate is then formed by determining the maximum with respect to the parameter a, i.e..

$$\& \, MAP = \max_{a} \{p(a)p(a)\}$$

(10.121)

where it is seen that the denominator $p(y)$ is not needed, since it does not depend on a. Thus the MAP estimate is computationally attractive. An intuitively satisfying method of deriving and understanding the MAP estimate is given by the following:

Suppose that the cost function in the conditional cost Eq. (10.80) is the uniform cost function of Eq. (10.86); then the average risk can be expressed as

∞

$\hat{a}(y)$-Δ/2

∞

R

$p(y)$

p(aly)da +

$p(aly)da$

dy

$\&()$+Δ/2

pô

(y)+Δ/2

-∞

$p(aly)$da dy

(10.122)

Ja(y)-Δ/2

Minimizing this equation is accomplished by selecting an estimate that maximizes the integral with respect to a. For small Δ, this integral is maximum for an estimate &MAP at which the *a posteriori* density p(ay) has a maximum. In the case of a unimodal *a posteriori* density (*i.e*. . only one peak), the estimate &MAP is the mode of p(a).[18] Note that & MAP can be found by setting the derivative of p(aly) to zero, *i.e.*,

a

$$-p(ay) = 0 \; \theta\alpha$$

(10.123)

and since the logarithm is a monotonic function, & MAP can also be determined by setting the derivative of the logarithm of p(ay) to zero, *i.e.*,

∂

$$\ln p(ay) = 0 \; \theta\alpha$$

(10.124)

These last two equations assume that the derivatives exist.

[17] See Problem 10.10.

Example 10.9

Using Example 10.7, it can be seen that the a *posteriori* pdf of Eq. (10.98) has a maximum when

$$\mu MAP =$$

$$y2w$$

=

$$\frac{\mathbf{B2y}+m102/m}{B2+o2/m} \quad (10.125)$$

Since the *a posteriori* pdf p(μÿ) is Gaussian, the mode and mean are identical, so that the MAP estimate is the same as the MSE estimate.

Example 10.10

This example illustrates a case where the MAP and MSE estimates are different. Let N, the number of vehicles per hour that arrive at an intersection, be Poisson distributed with parameter μ, *i.e.,*

$$p(N\mu) = \qquad N = 0, 1, 2\ldots$$

$$N!$$

18 Even if the pdf is not unimodal, the mode is still a reasonable choice for the MAP estimate.

(10.126)

The parameter μ is actually the average number of vehicles arriving per hour and is here assumed to be random with an exponential distribution in the parameter 2, *i.e.,*

$$p(\mu) =$$

$$\lambda\varepsilon\lambda\mu$$

$$0,$$

$$\mu \geq 0 \ \mu$$

$$< 0 \quad (10.127)$$

(Note that the average value of μ is 1/2.) The *a posteriori* distribution $p(\mu N)$ can be expressed as

where

$$p(\mu|N) = \frac{p(N\backslash\mu)p(\mu)}{P(N)}$$

p(N) =

═

[°

°

p(N\μ)p(u)*dμ*

∞

Substituting Eqs. (10.126) and (10.127) into Eq. (10.129) yields

∞

P(N)

═ [

(10.128)

(10.129)

λ

-λε-*μαμ*

∞

N = 0, 1, 2,...

(10.130)

N!

(2 + 1) N+1'

The *a posteriori* distribution *p*(*IN*) can then be written as

μNe-μ

λελμ

p(μ|N) =

N!

[2/(2+1)N+1]

∞

(λ + 1) N+1

N!

e ̄μ(2+1)

N = 0, 1, 2,..., μ≥0 (10.131)

The MSE estimate can be computed from the conditional mean equation, *i.e.*,

µ

PMS

=

√∘ ∘

(2 + 1)

N+1

N

N!

N+1

λ+1

An explicit expression for the conditional mean is

E{**N**} +1

E{µMS µ*}*

λ+1

Noting that *E*(Nu) u, the unconditional mean is expressed as

E{*E*{*λMS*\µ}} =

E{u} +1

+ *E*{*m*} λ+1

which shows that the estimate is biased.

(10.132)

(10.133)

(10.134)

The MAP estimate is obtained by finding the maximum of the *a posteriori* pdf *p*(u *N*), by differentiation, *i.e.*,

d [(a + 1) N+1

(2

du

N!

(2 + 1)N+1

N!

[NμÑ−1 e¯μ(a+1) + μ3[−(2 + 1)]e¯μ(2+1)]

Setting the derivative equal to zero yields the MAP estimate

N

μMAP

=

入+1

This estimate is different from the MSE estimate and is also biased, since

E(AMAP

}

E{*E*{**Nμ**}}

λ+1

E{*u*}

λ+1

(10.135)

(10.136)

(10.137)

These results indicate that the estimate depends on the cost function that has been selected. Both the MSE and MAP estimates are biased, although for large values of $E\{u\}$ they are approximately equal.

10.6.3 Absolute-Error Cost Function

Another estimate can be obtained by minimizing the conditional cost in Eq. (10.80) using the absolute-error cost function in Eq. (10.87). The average risk in this case can be expressed as

R

∞

-∞

P()

∞

la – & (y)\ *p*(*aly*) *da*
dỹ

pô
(ẏ)

(V
)

(a – â(ỷ))p(*α*|ỷ) da + + [
= *da*]

(a – *â* (y))*p*(*a*) *da dy* (10.138)

-∞

(V)

The average risk is minimized by minimizing the inner integral in the previous equation. Differ- entiating the inner integral with respect to *â* (v) by means of Leibnitz's rule [164] and setting the result to zero accomplishes the minimization, *i.e.*,

0 =

= ["_p(a\ÿ) da
— [® p(a\y) da

-∞

Rewriting this equation yields

L

p(a\ÿ*) da* = [®
p(a\ÿ) da

-∞

(10.139)

(10.140)

The resulting estimate for the absolute-error cost function is denoted by AE. This equation indicates that the absolute-error estimate is in fact the *median* of the a *posteriori* pdf *p*(*a*).19

[19] The median occurs when the cumulative distribution of a given ÿ is 1/2.

Note that if the *a posteriori* pdf *p*(*aly*) is symmetric, then the mean-square error estimate & *MS* and the absolute-error estimate & AE are equal, *i.e.*, MS = AE. For a unimodal and symmetric *a posteriori* pdf, the mean-square error estimate *âms* and the maximum *a posteriori* estimate â *MAP* are equal, *i.e.*, MS *MAP*; this result is observed in Example 10.9. For a Gaussian *a posteriori* pdf, which is both symmetric and unimodal, all three estimates are equal.

Example
10.11

Consider an *a posteriori* pdf given by

p(uly) =

y2

мe-*yн,*

μ> 0

0.

μ < 0

where μ is the parameter to be estimated. The absolute-error estimate is developed by

û

- yн

=

1

0

or

y2e-y

μ

v

y2

=

1

(10.141)

e ȳûAE (1+yû

Aε) ·

1

=

2

4

1

4

(10.142)

The MAP estimate can be computed for this case by differentiating the *a posteriori* pdf with respect to μ and setting the result to zero, *i.e.*,

1
0
or

d

(y2 не-yм) = 0 αμ

−ûye ӯû +e ӯû = 0

(10.143)

This equation can be solved for MAP, resulting in

$\mu MAP =$

y

(10.144)

The MSE estimate can be obtained by substituting the *a posteriori* pdf into Eq. (10.91), i.e.,

αμ

↑
MS

2

(10.145)

y

2μ 2 *12*

= [(3-3-3

]

The absolute-error estimate can be more readily compared with the MAP and mean- square error estimates by assuming the special case where yAE <1. As a result, the exponential can be expanded into a series where only the first two terms are used. Then Eq. (10.142) becomes

(1 − *yμ)* ≈
1/2

(10.146)

or

$$\mu AE = \frac{1}{\mathbf{2y}}$$

(10.147)

Thus, it is readily apparent that in this example all three estimates are different.

Example 10.12

The vehicle traffic example can be formulated for an absolute-error cost function. From the *a posteriori* pdf given by Eq. (10.131), the absolute-error estimate *MAE* is developed by the equation

μ

~A (2 +

1)N+1

$N!$

-

1

2

$N = 0, 1, 2, \ldots$

(10.148)

Let x = μ(+1) and $T = \mu(\lambda + 1)$. Then the preceding equation can be reexpressed

as

1

$N!$

√ √ x " e

"dx = {}

2

Using [52] allows Eq. (10.149) to be expressed in series form as

ex

+

$N!$

Σ

(-1)*

(-1)k+1

NkxN-k

-

k=1

where

(10.149)

(10.150)

$NkN(N-1)\ (Nk+1)$

Equation (10.150) can be simplified to

e^{-1}

N

$N!$

$TN +$

ΣN_TN-k

$(+ \Sigma) -$

2

(10.151)

(10.152)

Further simplification is not readily accomplished without detailed numerical computation. A numerical example can be obtained by letting N = 2. In this case, the foregoing equation is

e^{-T}

$2\ (T2 + 2T$

$+ 2)$

$=$

$\{\}$

(10.153)

A numerical solution results in T≈ 2.67 or

MAE ≈

2.67

(+1)

(10.154)

10.7 MINIMAX ESTIMATION

The average risk in Bayes estimation theory is given by Eq. (10.79). An alternative form of the average risk can be expressed as

$$R= \text{LL} \; C(a, \hat{a}(y))p(a)p(ya)dad\tilde{y} \qquad (10.155)$$

In minimax estimation, an *a priori* distribution $p(a)$ is selected that maximizes the Bayes risk given by Eq. (10.155); the resulting estimate, formed by minimizing the maximum risk, is a minimax estimate &MMX with an average risk RMMX given by 20

$$RMMX = \; =\text{PL} \; C(\alpha, \hat{a}_{MMX}(y))p(a)p(ya)dady \quad {}^{\infty} \; {}_{-\infty} \qquad (10.156)$$

By definition, R ≤ RMMX no matter what the true *a priori* pdf of the random variable a is postulated to be.21

Example 10.13

In Example 10.7 the estimate of the mean was found to be

$$\mu MS \quad \frac{B2y+m102/}{m\, B2 + o2/m}$$

with an average minimum risk given by

$$Rmin \quad {}^{m} \quad \frac{B2}{B2 + o2/m} \qquad (10.157)$$

(10.158)

Recall that the true mean was assumed to be Gaussian with mean m1 and variance *ß2*. Suppose that the *a priori* pdf of the true mean is unknown *i.e.*, it has a variance that approaches infinity (82 → ∞). In this case the estimate MS→y and the average risk Rmino2/*m*. In other words, the minimax estimate MMX = **y** (the sample mean) and RMMX = 02/*m* (the minimax risk). Thus the average risk obtained by maximizing Rmin is o2/*m* no matter what the actual a *priori* distribution of μ happens to be, and

σ

[20] Weber [160] indicates that minimizing the maximum risk and finding the maximum of the minimum risk are equivalent.

[21] See also [58] and [57].

the average risk is always less than *RMMX* = o2/*m*. This example points out that if the Bayes solution has constant risk, then the estimate is a minimax estimate.

It should be noted that explicit minimax estimates are not generally easy to find. Further examples are provided in [83].

10.8 MAXIMUM-LIKELIHOOD ESTIMATION

Bayes estimation requires knowledge of the *a posteriori* distribution *p*(a), which, in turn, requires knowledge of the *a priori* distribution *p*(a) and the likelihood distribution *p*(a). On the other hand, ML estimation is based on maximizing the likelihood distribution and thus requires no knowledge of either the *a priori* pdf or any cost function. It should be noted that Bayes estimation may result in more accurate estimates if the statistical distributions and cost functions are known, but the Bayes estimate can lead to different inferences from the same observed data if these items are postulated and possibly inaccurate. In contrast, ML estimation is often criticized because it ignores *a priori* information. One nice feature of ML estimates is that they are typically easier to compute and often yield the best estimate of the parameter.

In this section it is assumed that the *m* observations of the random variable **y** with a parameter a are made. Each observation **y**; has a pdf *p(y;* la) and, since the *m* observations are independent, the joint pdf pyla), known as the likelihood function, can be expressed as

$$p(\tilde{y}|\alpha) = [[_{i=1}^{m} p(y;\backslash\alpha) \qquad (10.159)$$

A maximum-likelihood estimate &ML of the parameter **a** is then determined by finding the maximum of the likelihood function *p(ya)*. Computationally, the maximum is obtained by setting the derivative of the likelihood function to zero, i.e.,

$$\frac{a}{\theta\alpha} p(ya) = 0 \tag{10.160}$$

Since the logarithm is a monotonic function, a simpler procedure is often used where the maximum of the logarithm of the likelihood (referred to as log-likelihood) is computed, *i.e.*.

$$\frac{a}{\theta\alpha} \text{In } p(y|a) = 0 \tag{10.161}$$

Again, these two procedures are used only if the derivatives exist.

The difference between ML and MAP estimation can be viewed intuitively by examining the relationship

$$p(ay) = \frac{p(a)p(a)}{p(y)} \tag{10.162}$$

In MAP estimation, the maximum of the *a posteriori* pdf $p(aly)$ is desired and can be obtained from the derivative of the logarithm of *p(aly)*, i.e.,

$$\frac{\partial}{\theta\alpha} \text{In } p(aly) = \frac{\partial}{\theta\alpha} \text{In } p(ya) + \frac{a}{\theta\alpha} \text{In } p(a) - \frac{a}{\theta\alpha} \text{In } p() \tag{10.163}$$

Note that the derivative of the In $p(y)$ is zero, since $p(v)$ does not depend on a. If the *a priori* pdf $p(a)$ has a wide dispersion (e.g., it is uniformly distributed over a wide range), then the second term in the preceding relationship is nearly zero. As a result, the estimate & that maximizes *p(*aly) (the MAP estimate) is nearly the same as the estimate that maximizes *p(*ya) (the ML estimate). Thus, an important advantage of an ML estimate is that it can be obtained when no information about the *a priori* pdf is available. It should be realized, however, that with knowledge of the *a priori* pdf, an ML estimate will be poorer than a MAP estimate.

Other important properties of ML estimation are as follows:

1. If an efficient estimate exists, it is an ML estimate.

2. An ML estimate is invariant; that is, if *g*(a) is invertible for all a, then the ML estimate of *g*(a), expressed as $\hat{g}ML$, is ĝML = g(*ML*). Invariance of an ML estimate allows unknown parameters to be replaced by their ML estimate and leads to a generalized likelihood ratio test, as indicated on page 200 of [69].

3. The ML estimate is asymptotically efficient, best asymptotically normal (BAN), consis- tent, and is a function of sufficient statistics for general regularity conditions *(i.e.*, no singularities) of the pdfs *p(y;* la). See [93], [164], [57], and [152].

Example 10.14

This example is a continuation of Example 10.7, where the mean ***u*** is to be estimated for a Gaussian distribution with known variance o2 from the independent observations yi,**...,** ym, where each observable has pdf

$$P(yil\mu) = \quad 1 \quad \exp \quad (Yi - \mu)2 \quad 202 \qquad i = 1,..., \quad ..., m \qquad (10.164)$$

The likelihood function is then

$$py\mu) = \quad 1 \quad (2\pi\sigma2)m/2 \quad \exp \quad \Big(\quad m \quad 1 \quad 202 \quad \Sigma O \quad K \quad \mu)2 \qquad (10.165) \quad k=1$$

and the log-likelihood obtained from the latter equation is

$$\ln p(\tilde{y}|\mu) = \overset{m}{—} "$$

ln(27σ2)

2

m

1

202

Σοκ - μ)

k=1

(10.166)

Differentiating the preceding equation with respect to μ and setting the result to zero yields

-

1

6

m

Σ(yk - μ) =

0

k=1

(10.167)

or

μM

L

Σ

Yk=**y**

m

k-1

(10.168)

Therefore, the ML estimate in this case is the sample mean.

Properties of the sample mean have been shown to be unbiased, consistent, BAN, a sufficient statistic, and efficient (i.e., minimum variance among unbiased estimates). Therefore, when the ML estimate is the sample mean, it exhibits desirable properties of a good estimate. In other instances, the ML estimate may not have all of these properties. For example, the ML estimate is often biased, as indicated by further examples in this section.

Example 10.15

This example is a continuation of Example 10.10, on the vehicle traffic at an intersection. The likelihood function in this example is *p(N*|μ) =

N = 0, 1, ..., 2

N!

(10.169)

The log-likelihood can then be written as

In *p(N*) *N* In μ-μ- In *N*!

(10.170)

Differentiating the preceding equation with respect to μ and setting the result to zero leads to

or

N

0

μ

μMI. = **N**

(10.171)

(10.172)

This estimate is unbiased, since the *E*{ML} = *E*{**N**}, where the average value of the Poisson distribution is *E*{**N**} = *μ*.

Example 10.16

This example assumes that the observable **y1**, ..., **Ym** are each independent and have a Gaussian distribution, where both the mean μ and variance o2 are to be estimated. The likelihood function is then

m

pylμ, o2)=

1

(2702)m/2

exp

1 202

Σ(yk -

μ)2

(10.173)

k=1

with log-likelihood

In *p(*ỹ|μ, σ2) = –TM

In(27 o2)

m

1

2

202

Σ(yk -

μ)2

(10.174)

k=1

Letting so = o2, the ML estimates μML and SOML are obtained by computing the follow- ing derivatives:

in

a

1

In pyμ, so) =

Σoκ - μ)

(10.175)

θμ

SO

k=1

in

a

m

1

In pμ, so)

"

∂ѕo

+ Σ(yk - μ)2

(10.176)

2.50 256

k=1

Setting Eq. (10.175) to zero yields the ML estimate of the mean, ML = y. Using this result in Eq. (10.176) and setting it to zero yields the variance estimate, *i.e.*,

m

,1

1

Σ(yk - y)2 =

0

k=1

+ 250 23

(10.177)

or

m

SOM

L

k=1

(10.178)

The ML estimate of the mean is unbiased, but the ML estimate of the variance, which is the sample variance defined in Section 10.5.1, is biased. Both estimates are sufficient statistics, but the sample variance modified by the factor to create s = SOML

i

s only asymptotically efficient.2

22

m

m-1

m

m

An important property of the asymptotic distribution of an ML estimate is that it is normally distributed. If $\hat{a}$ is an ML estimate of a, with true mean a, the pdf of $\hat{a}$ is Gaussian with mean

a2 In pa) is the Fisher information.

a and variance $F1$(a), where F

$(a) = -E$ {

Example 10.17

}

Assume the pdf for a random vector y is given by

$p(yi, \alpha) = aya$-1,

$0 < y < 1$, $i = 1, \ldots, m$

(10.179)

and 0 otherwise. Then the joint density function is

m

$p(\tilde{y}, \alpha) = a$" [ ]

ya−1

i=1

(10.180)

and the natural logarithm is

In $p(\ddot{y}$, a) = m Ina + (a− 1) In y;

(10.181)

i=1

[22] See Sections 10.5.4 and 10.5.6.

Forming the partial derivative al p.) and setting the result to zero yields

θα

m

+In y; =0

α

i=1

or, solving for the estimate, we have

m

â

In yi

(10.182)

(10.183)

The Fisher information can then be computed as

m

22 In *p(y*, a) *E* = "

F(a) =

-*E*

да2

(10.184)

From these results, we know that â is Gaussian distributed with mean a and variance a2*/m.*

10.9 COMPARISON OF ESTIMATORS OF PARAMETERS

In this section several comparisons will be made for estimators of parameters, using the techniques described earlier. A summary of results for the problem of estimating the mean u from the observations **y1**, ..., **ym**, which are each independent Gaussian with known variance o2, is provided along with specific estimation properties. The pdf *p(*u) and a *priori* pdf p(μ) are given as

1

pyμ) =

m

1

=

(2702)m/2

exp Σoκ - *μ)2*

202

k=1

1

$p(\mu) =$

1

exp

$2\pi\beta$

2

2B2

(μ.

the MSE, MAP, and ML estimates are, respectively,

μMS =

B2y+m102/m

B2 + o2/*m*

(10.185)

μMAP=

MAP

B2y+m102/

m

-

μMS

B2 +

o2/*m*

μML = **y**

(10.186)

The first two estimates are identical and biased; the third is unbiased and efficient. For large *m* or large B2, MS≈ ML So that all three estimates are approximately equal. Large B2 is equivalent to an *a priori* pdf with a wide dispersion, so that little statistical information is gained

from knowledge of the *a priori* pdf. In this case, the MSE and MAP estimates reduce to the ML estimate, which does not require an *a priori* pdf.

Consider the vehicle traffic example with

and

$p(N\backslash\mu) =$

e-*H*

N!

$N = 0, 1, \ldots$

(10.187)

$p(\mu)$

=

{

λελμ

0,

μ < 0

(10.188)

The MSE, MAP, and ML estimates in this case are, respectively,

N+1

PM

S

=

2+1

N

λ+1

μ MAP

=

MAP

AML = **N**

Again, the MSE and MAP estimates are biased and the ML estimate is not.

An example where all three estimates are different is as follows:

Example 10.18

ni

(10.189)

The independent observables are obtained from **y1** = for *i* = 1,..., *L*, where **a** is a random variable parameter to be estimated and the pdfs for n; and a are, respectively,

λελη , ni > 0

p(ni) =

{

0,

ni ≤ 0

and

p(a)

= {

ae-aa,

0,

α ΣO

a<0

(10.190)

The MSE estimate is obtained from Eq. (10.94), *i.e.*,

Soap*(a)ae*-aα *da*

&

MS

s

Sopyla*)ae-aada*

αα

(10.191)

The pdf *p(y; a*) can be computed by a transformation of random variables (with a held fixed) and calculation of the cdf, *i.e.*,

Pr[y; < yila] = Pr (TM <yi|a) =

Pr[n; <y;α|α]

για

α

λe-λni

dni

= 1-e-*hayi*

v

λ

Yi > 0

(10.192)

The pdf *p(ya*) is then obtained by differentiating the cdf:

*p(*yilα) =

==

{

λα-λαγι 0,

Yi > 0 *Yi* < 0

(10.193)

The joint pdf $p(ya)$ is the product of the individual pdfs, *i.e.*,

$$p(a) = (\lambda a)1 \exp$$

$$-\lambda\alpha\ \Sigma$$

$$i=1$$

(10.194)

Identifying the sample mean, y, as **y** = y, the preceding equation becomes

$$p(\tilde{y}|\alpha) =$$

$$(\lambda\alpha)1e-haLy$$

(10.195)

Substituting Eq. (10.195) in Eq. (10.191) yields

$$a$$

$$So\ \alpha\ (\lambda\alpha)1e-\alpha\ \ Lyae-aa$$

$$da$$

$$\hat{a}\ MS$$

(10.196)

$$So\ (2\alpha)-e-\lambda\alpha$$

$$La\alpha e-aada$$

From [52], it is known that

$$f$$

$$x''\ e-cx\ dx =$$

$$n!$$

$$ch+1'$$

$$c> 0$$

(10.197)

Using this result allows the integrals in Eq. (10.196) to be evaluated as

$$\hat{a}ms$$

$$1\ (L+1)!/(L\tilde{y}+a/\lambda)L+2$$

$$\lambda$$

$$L!/(LT + a/\lambda)\ 4+1$$

or

$$L+1$$

$$\hat{a}\ MS$$

$$\lambda\ \mathbf{Ly} + a$$

(10.198)

The MAP estimate can be computed by finding the maximum of $p(a)$ or equivalently by finding the maximum of In p(a)p(a). The maximum is then obtained from

∂

a

In p(la) + θα

In p(a) = 0

θα

Substituting Eq. (10.195) and (10.190) in (10.199) yields

or

(10.199)

∂

-λα Ly

θα

~ In [(20) e-al)] += In

[ae] = 0

a

θα

αα

(10.200)

∂

[L In λa - λa $\boldsymbol{Ly}$ + In a − aa]

= 0

(10.201)

$\theta\alpha$

Computing the derivative and solving for a results in

L

& MAP

=

$\lambda Ly+a$

(10.202)

The ML estimate is obtained by computing the maximum of p(a). Using the log-likelihood, the ML estimate is determined from

or

a

$[L \text{ In}(\lambda a) - \lambda a \text{ Ly}] = 0$

θα

(10.203)

&

ML

-

λγ

(10.204)

In summary, the MSE, MAP and ML estimates for this example are, respectively,

$L+1$

&

MS

$\lambda \text{ Ly} + a$

L

&

MAP

-

$\lambda \text{ Ly} + a$

&

ML

-

1

Xy

(10.205)

For small *a,* the MAP and ML estimates approach each other, and for large *L*, all three estimates are nearly the same. All three estimates are biased.23

Example 10.19

Another example in which all three estimates can be determined is provided here. Consider a gamma distribution with parameters *u* and λ and a pdf given by

p(yu, λ):

I

Qyju-

1

-

$\Gamma(u)$

(u) de-y 0,

$y \geq 0$ y

<0

(10.206)

For integer *u* = n, from [75], the gamma distribution specializes to an *n*-stage Erlangian distribution. Such a distribution arises in queueing theory where a customer enters a service facility and must proceed through *r* stages before exiting. As the old customer exits the rth stage and leaves the service facility, a new customer may enter. For example, in a two-stage Erlangian case, Eq. (10.206) becomes

$$P(y|2) = \begin{cases} \lambda 2ye\text{-}\lambda y, & y\geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad (10.207)$$

[23] **See** Problem 10.15.

The pdf in Eq. (10.207) is obtained by determining the sum **y** of two independent random variables y1 and y2, each having the same pdf

$$\lambda\varepsilon\lambda\varepsilon\iota\ P(y;2) = \begin{cases} & Yi \geq 0 \\ 0, & Yi < 0,\ i = 1,2 \end{cases} \qquad (10.208)$$

This equation is the pdf of a random variable that represents the service time at stage one or two. Thus the average service time at stages one or two is *E*{y} = *E*{y2} = 1/λ and the variance is V{y1} = *V* {y2} = 1/22. As the customer proceeds from the first stage to the second, the total amount of time the customer spends in the facility is a random variable **y** = **y1**+ y2 with a pdf given by Eq. (10.207). The average time the customer spends in the service facility is then *E*{**y**} = 2/2 with a variance V{**y**} = 2/22.

Now suppose that the parameter **λ** is itself a random variable with parameter *a* and pdf given by

$$p(\text{x}) = \begin{cases} \text{a}e\text{-}a\lambda, & \lambda\geq 0 \\ 0, & \text{otherwise} \end{cases} \qquad (10.209)$$

We can now determine the MSE, MAP, and ML estimates for the service-rate parameter λ as well as the average risk and average values of the estimates.

The *a posteriori* pdf *p*(2|y) is computed from Bayes' theorem as

p(λly) =

p(yλ)p(λ)

p(y)

(10.210)

where, from the theorem of total probability,

∞

p(y) = √

_~_p(y\^)p(*x*)*d*

}

-∞

Using Eqs. (10.207) and (10.209) in Eq. (10.211) results in

∞

P(v) = √   √®

22 *ye*˥*3* ae

¯aλdλ =

The *a posteriori* pdf then becomes

(10.211)

2ay (a + y)

3'

y>0

(10.212)

2

2

λ2

ye-λyae-ax

αραλ

p(λy) =

(a +

y)31⁄22e¬λ(*a*+*y*)

2ay

$\lambda > 0$

2

(10.213)

(a+y)

3

The MSE estimate can now be calculated as

λms = [° °

λp(^\y)dλ = [°

p∞ λ3 (*a* +

y)3e−λ(a+y)

-dλ

(10.214)

2

Section 10.9 Comparison of Estimators of Parameters

or

ÂM

S

3

a + y

(10.215)

The MAP estimate is obtained by finding the maximum of In *p(*y|2*)p(*2) from

a

[In (22 ye-*ae*-aλ)]

= 0

θλ

Evaluating this expression leads to

or

10

(10.216)

λ

212

-y-a = 0

(10.217)

ÂM
AP

2

a + y

The ML estimate is computed by finding the maximum of In *p(*y|λ) from

a

[In (22 ye¯y2)]
= 0

θλ

After evaluation, this results in

2

ÂM
L

y

(10.218)

(10.219)

(10.220)

The average values of each estimate, with λ held content, can be found as follows. For the ML estimate,

E {λML}
= *E*

22ye¯y^*dy* =
22

The MAP estimate has **an** average value (for a constant λ) given by

∞ 2

a + y

E (MAP) = *E*{2} = [

7, 22 *ye*="" *dy*

a+

Using the substitution x = *a* + *y*, this expression can be calculated to be

(10.221)

(10.222)

E{ÂMAP} = 22 - 222aeaλ
E1 *(aλ)*

(10.223)

where *E1(u)* = $\overset{u}{x}$-le-**dx* is the exponential integral, defined in [1].

Using the same procedure, the average value of the MSE estimate, for a constant 2, can be written as

E(AMS) =
E

3

=32-322aeaλ *E1* (aλ)

(10.224)

Since none of the estimates results in *E*{â} = 2, all three estimates are biased. As a numerical example, assume that *a* = 1 and that λ is fixed at λ = 1 as well. In this case, *E1*(1) = 0.219 and

E{ÂMS} =
1.214

E{ÂMAP} =
0.809

E{λML} =
2

(10.225)

For this example the ML estimate has the highest average service rate, whereas the MAP estimate has the lowest.

Using MATLAB, plots shown in Figure 10.6 are obtained for the expected value of the ML, MS, and MAP estimates for $a = 1$. The MATLAB routines used in the computation are given in the website www.prenhall.com/schonhoff, lam.m is the primary routine, which calls e1.m and elint.m.

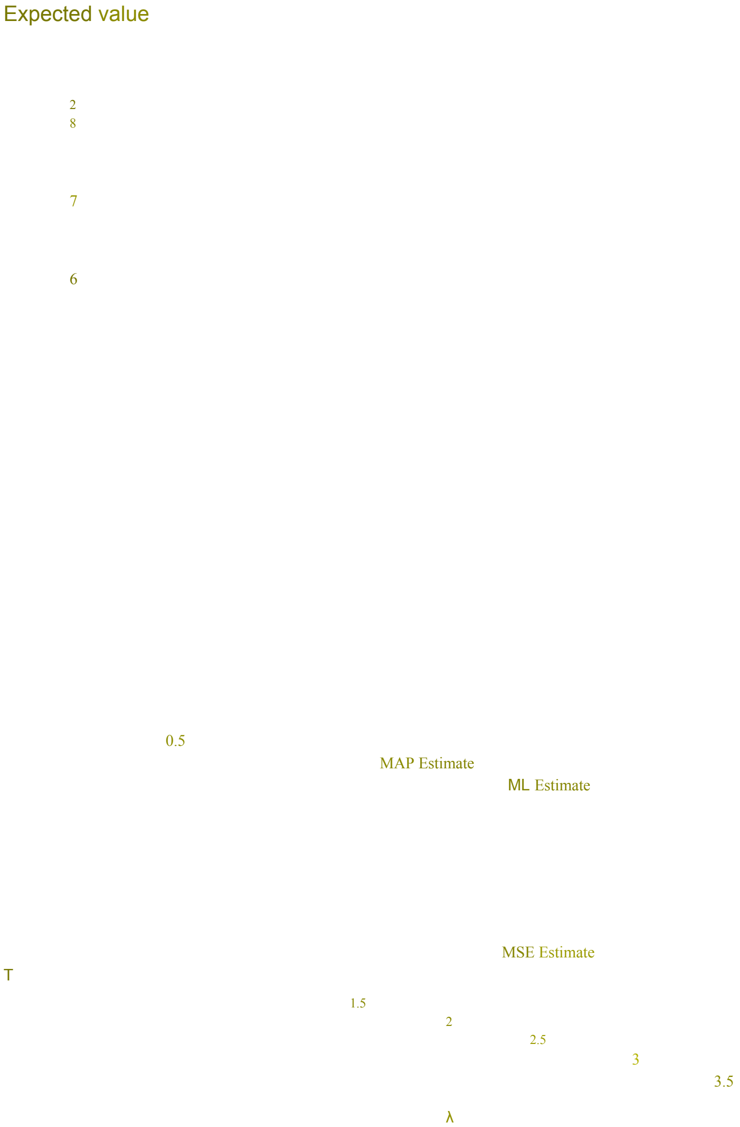


Figure 10.6. Expected value of ML, MS, and MAP estimates of service rate

We can now determine the average risk associated with each service-rate estimate. The average risk using the MSE cost can be computed with

$$R$$

$$=$$

[__ [__

Q

(2- -

^)2p(^\y)p(y)dλdy

Substituting Eqs. (10.207) and (10.209) into this equation results in

∞

R

(2-2)2aλ2ye-λ(a+y) dλdy

For the ML estimate, the average risk *RML*, using an MSE cost, is

2 2

RML =

λ

aλ2 ye-λ(a+y)

dλdy

Expanding the square yields the expression

RML

= (1

(10.226)

(10.227)

(10.228)

∞

4

λ4e−λ(a+y) dλ

−4

23

[©

22e=λ(a+y)

dx] dy

1

x

]

7

(10.229)

Using the general expression for integer *n* and constant *c*,

ecx

n

2-1

[xe dx = x" "

– " fx - *1* e

"dx

where in the special case

C

xedi Jxe

xecx dx =

ecx

c2

(cx - 1)

allows the evaluation of the individual terms in RML, so that

x

RML=24a

y

—

x

1

(a + yjsdy - 240 f (a

+ yids+8a

y)4

(10.230)

(10.231)

∞

1

y*(a* +

vzdy

(10.232)

These terms can be evaluated using the identity for integers m and n:

∞

xm-1

-dx

(1 + x)m+n

Γ(m)Γ(n)

Γ($m + n$)

(10.233)

Here, (n) is the gamma function and for integer n is Γ(n) = (n − 1)!. Completing the computations in Eq. (10.232) results in

RML

2

42

8

8

a2 + 2 Γ

(0)

(10.234)

However, (e) approaches infinity as approaches zero, so that the RML is infinite.

In the MSE case, the average risk can be written as

RMS = a

λ

a + y

Following the same procedure outlined above yields

RMS

=a

2

3

yλ2e-λ(a+y)

$d\lambda dy$

(10.235)

[

0

0

λ3e-λ(a+y)

dλ

a

x]

λ4e-λ(a+y)

dλ

9y +

(a + y)

2

y

(a + y)5

dy

S

= 6a

For the MAP case,

RMAP = a

a fo

for (

∞

λ

8

6y

a + y Jo

x2 e-λ(a+y)

dx dy

1

2a2

2

(10.236)

y

a + y

yλ2e-λ(a+y)

dλdy

2

(10.237)

8a [(a +

sdy = 32/3

0 (a + y)5

As expected, the average risk is smaller for the MSE estimate than for the MAP estimate.

Example 10.20

A final example, which is more representative of communications theory, is to consider the reception of a signal embedded in noise. In this case, the *L* independent observables or received-signal samples y, are expressed as the sum of a random signal **s** and noise samples n;, *i.e.,*

Yi = **s**+ni,

i = 1,..., *L*

(10.238)

It is further assumed that the noise is Gaussian with zero mean and variance o2 and that the signal is uniform in the interval (0, a), so that the pdfs can be written as

p(s) =

0,

0 < s <a

otherwise

1

p(ni) =

-n2/20

2

2πσ2

(10.239)

For a constant s, y; is Gaussian with means and variance o2, thus allowing the pdf *p(*s) to be written as

pys)

1

1

P(515) = (2014/2

exp(-2,7 20,-3)2)

(2πσ2

202

(10.240)

Computing the maximum-likelihood estimate from *p(*s) = 0 results in

SML = **y**

@ In

as

(10.241)

The MAP estimate is obtained from In *p(*y|*s)p*(s) = 0 and is therefore the same as the ML estimate, *i.e.*,

SMAP SML = **y**

The MSE estimate is obtained from Eq. (10.94), *i.e.*,

s

So å (2703)4/2 exp

(−232 Σ÷=1(vi − s)2) ds

SMS

So a(2702)1/2 exp (−262

Σi=1(vì − s)2) ds

This expression can be rewritten by expanding the sum in the exponential, *i.e.*,

L

Σ(yi-s)2=Σy? - 2s Σyi +Σ

i=1

i=1

Σ+Σ

i=1

Σy-2s *Ly* + *Ls2*

i=1

Using this equation in Eq. (10.243) leads to

SM
S

Jos exp(-4)*ds*

(st

y

fo exp (347 -

224) ds

02

20

i=1

Completing the square in the exponential allows Eq. (10.245) to be written as

(10.242)

(10.243)

(10.244)

(10.245)

202

L

Jos exp(-2(-9)2)

ds

SMS

=

So exp(-2 (sy)2) *ds*

202

After considerable algebra this equation takes the form24

SM

S

√2/10 (exp(-121)

- exp(-

π ι

-y)

L(*a*-y)2

20-

erf ((√) + erf

(√)

σ

σ

:)

$$\Big) \qquad +\tilde{\mathbf{y}} \qquad (10.246)$$

$$(10.247)$$

Thus the MSE estimate is biased, whereas the ML and MAP estimates are unbiased. The term summed with **y** is the bias and can be numerically evaluated using the MATLAB

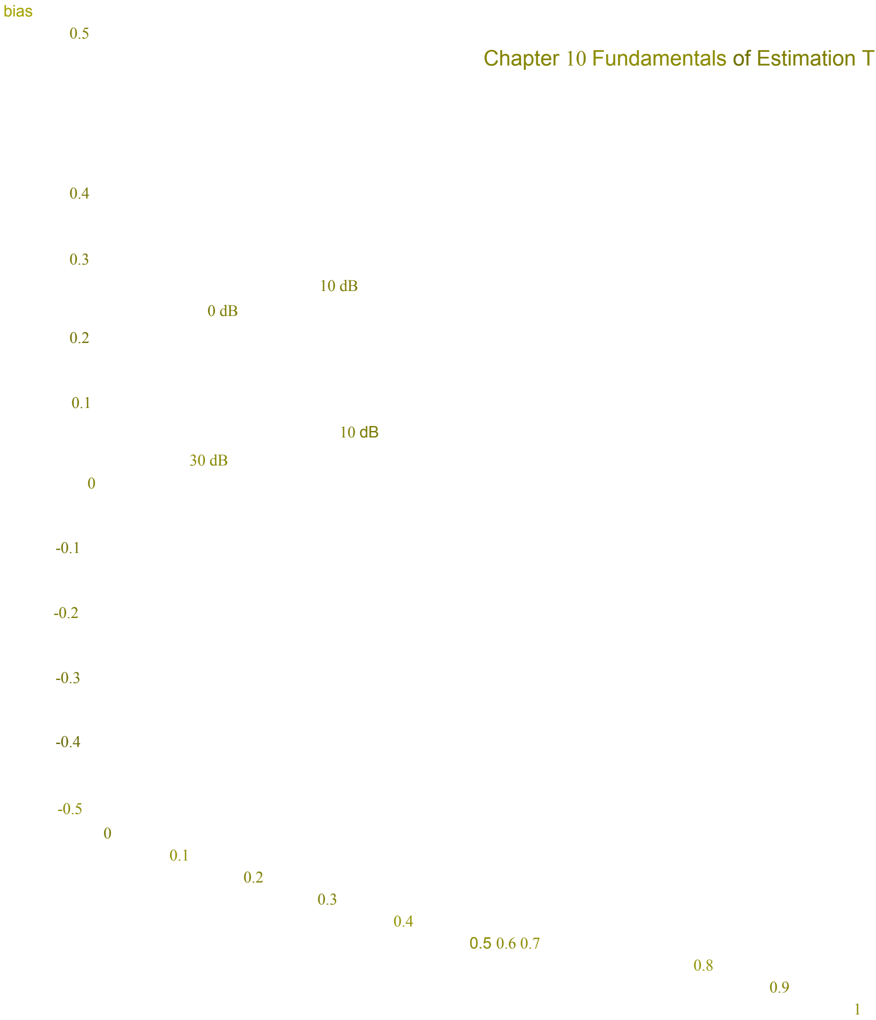


Figure 10.7. Bias as a function of **y**

[24] See Problem 10.16.

routine bias.m given in the website www.prenhall.com/schonhoff. The quality of the estimate is a function of the parameter $L/(202)$, which is defined using decibels as

$$8dB = 10 \log \frac{L}{202} \qquad (10.248)$$

Figure 10.7 illustrates the bias parametric in gaB. From the figure it can be seen that as gdB increases, the bias term approaches zero.

10.10 BIBLIOGRAPHICAL NOTES

The material in this chapter is based on numerous classical statistics references, e.g., [86], [58], [33], and [35], as well as communications engineering references such as [152], [164], [90], [93], [57], and [69]. Classical statistics references typically describe properties of good estimators, whereas communication engineering references focus more heavily on Bayes, MSE, MAP, and ML estimation. Since many systems operate in the presence of Gaussian noise, estimators which are Gaussian derived are emphasized.