which measures the *average* curvature of the log-likelihood function. The expectation is taken with respect to $p(x[0];A)$, resulting in a function of A only. The expectation acknowledges the fact that the likelihood function, which depends on $x[0]$, is itself a random variable. The larger the quantity in (3.5), the smaller the variance of the estimator.

3.4 Cramer-Rao Lower Bound

We are now ready to state the CRLB theorem.

Theorem 3.1 (Cramer-Rao Lower Bound - Scalar Parameter) *It is assumed that the PDF $p(\mathbf{x};\theta)$ satisfies the "regularity" condition*

$$E\left[\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}\right] = 0 \qquad \textit{for all } \theta$$

where the expectation is taken with respect to $p(\mathbf{x};\theta)$. Then, the variance of any unbiased estimator $\hat{\theta}$ must satisfy

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{-E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right]} \tag{3.6}$$

where the derivative is evaluated at the true value of θ and the expectation is taken with respect to $p(\mathbf{x};\theta)$. Furthermore, an unbiased estimator may be found that attains the bound for all θ if and only if

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} = I(\theta)(g(\mathbf{x}) - \theta) \tag{3.7}$$

for some functions g and I. That estimator, which is the MVU estimator, is $\hat{\theta} = g(\mathbf{x})$, and the minimum variance is $1/I(\theta)$.

The expectation in (3.6) is explicitly given by

$$E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right] = \int \frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2} p(\mathbf{x};\theta)\, d\mathbf{x}$$

since the second derivative is a random variable dependent on $\mathbf{x}$. Also, the bound will depend on θ in general, so that it is displayed as in Figure 2.5 (dashed curve). An example of a PDF that does not satisfy the regularity condition is given in Problem 3.1. For a proof of the theorem see Appendix 3A.

Some examples are now given to illustrate the evaluation of the CRLB.

Example 3.2 - CRLB for Example 3.1

For Example 3.1 we see that from (3.3) and (3.6)

$$\mathrm{var}(\hat{A}) \geq \sigma^2 \qquad \text{for all } A.$$

Thus, no unbiased estimator can exist whose variance is lower than σ^2 for even a single value of A. But in fact we know that if $\hat{A} = x[0]$, then $\mathrm{var}(\hat{A}) = \sigma^2$ for all A. Since $x[0]$ is unbiased and attains the CRLB, it must therefore be the MVU estimator. Had we been unable to guess that $x[0]$ would be a good estimator, we could have used (3.7). From (3.2) and (3.7) we make the identification

$$\begin{aligned} \theta &= A \\ I(\theta) &= \frac{1}{\sigma^2} \\ g(x[0]) &= x[0] \end{aligned}$$

so that (3.7) is satisfied. Hence, $\hat{A} = g(x[0]) = x[0]$ is the MVU estimator. Also, note that $\mathrm{var}(\hat{A}) = \sigma^2 = 1/I(\theta)$, so that according to (3.6) we must have

$$I(\theta) = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right].$$

We will return to this after the next example. See also Problem 3.2 for a generalization to the non-Gaussian case. ◇

Example 3.3 - DC Level in White Gaussian Noise

Generalizing Example 3.1, consider the multiple observations

$$x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $w[n]$ is WGN with variance σ^2. To determine the CRLB for A

$$\begin{aligned} p(\mathbf{x};A) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n]-A)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]. \end{aligned}$$

Taking the first derivative

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x};A)}{\partial A} &= \frac{\partial}{\partial A}\left[-\ln[(2\pi\sigma^2)^{\frac{N}{2}}] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \\ &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) \\ &= \frac{N}{\sigma^2}(\bar{x} - A) \end{aligned} \tag{3.8}$$

where $\bar{x}$ is the sample mean. Differentiating again

$$\frac{\partial^2 \ln p(\mathbf{x};A)}{\partial A^2} = -\frac{N}{\sigma^2}$$

and noting that the second derivative is a constant, we have from (3.6)

$$\mathrm{var}(\hat{A}) \geq \frac{\sigma^2}{N} \tag{3.9}$$

as the CRLB. Also, by comparing (3.7) and (3.8) we see that the sample mean estimator attains the bound and must therefore be the MVU estimator. Also, once again the minimum variance is given by the reciprocal of the constant N/σ^2 in (3.8). (See also Problems 3.3–3.5 for variations on this example.) ◇

We now prove that when the CRLB is attained

$$\mathrm{var}(\hat{\theta}) = \frac{1}{I(\theta)}$$

where

$$I(\theta) = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right].$$

From (3.6) and (3.7)

$$\mathrm{var}(\hat{\theta}) = \frac{1}{-E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right]}$$

and

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} = I(\theta)(\hat{\theta}-\theta).$$

Differentiating the latter produces

$$\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2} = \frac{\partial I(\theta)}{\partial \theta}(\hat{\theta}-\theta) - I(\theta)$$

and taking the negative expected value yields

$$\begin{aligned}-E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right] &= -\frac{\partial I(\theta)}{\partial \theta}[E(\hat{\theta})-\theta] + I(\theta)\\ &= I(\theta)\end{aligned}$$

and therefore

$$\mathrm{var}(\hat{\theta}) = \frac{1}{I(\theta)}. \tag{3.10}$$

In the next example we will see that the CRLB is not always satisfied.

Example 3.4 - Phase Estimation

Assume that we wish to estimate the phase ϕ of a sinusoid embedded in WGN or

$$x[n] = A\cos(2\pi f_0 n + \phi) + w[n] \qquad n = 0,1,\ldots,N-1.$$

The amplitude A and frequency f_0 are assumed known (see Example 3.14 for the case when they are unknown). The PDF is

$$p(\mathbf{x};\phi) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}[x[n]-A\cos(2\pi f_0 n+\phi)]^2\right\}.$$

Differentiating the log-likelihood function produces

$$\begin{aligned}\frac{\partial \ln p(\mathbf{x};\phi)}{\partial \phi} &= -\frac{1}{\sigma^2}\sum_{n=0}^{N-1}[x[n]-A\cos(2\pi f_0 n+\phi)]A\sin(2\pi f_0 n+\phi)\\ &= -\frac{A}{\sigma^2}\sum_{n=0}^{N-1}[x[n]\sin(2\pi f_0 n+\phi) - \frac{A}{2}\sin(4\pi f_0 n+2\phi)]\end{aligned}$$

and

$$\frac{\partial^2 \ln p(\mathbf{x};\phi)}{\partial \phi^2} = -\frac{A}{\sigma^2}\sum_{n=0}^{N-1}[x[n]\cos(2\pi f_0 n+\phi) - A\cos(4\pi f_0 n+2\phi)].$$

Upon taking the negative expected value we have

$$\begin{aligned}-E\left[\frac{\partial^2 \ln p(\mathbf{x};\phi)}{\partial \phi^2}\right] &= \frac{A}{\sigma^2}\sum_{n=0}^{N-1}[A\cos^2(2\pi f_0 n+\phi) - A\cos(4\pi f_0 n+2\phi)]\\ &= \frac{A^2}{\sigma^2}\sum_{n=0}^{N-1}\left[\frac{1}{2}+\frac{1}{2}\cos(4\pi f_0 n+2\phi) - \cos(4\pi f_0 n+2\phi)\right]\\ &\approx \frac{NA^2}{2\sigma^2}\end{aligned}$$

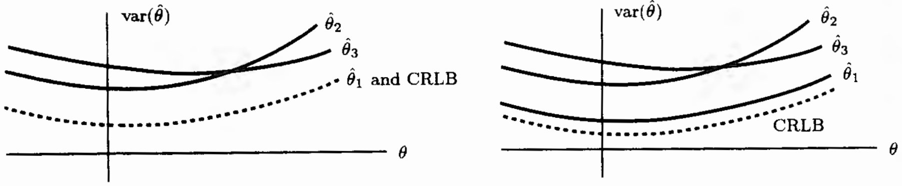
since

$$\frac{1}{N}\sum_{n=0}^{N-1}\cos(4\pi f_0 n+2\phi) \approx 0$$

for f_0 not near 0 or 1/2 (see Problem 3.7). Therefore,

$$\mathrm{var}(\hat{\phi}) \geq \frac{2\sigma^2}{NA^2}.$$

In this example the condition for the bound to hold is not satisfied. Hence, a phase estimator does not exist which is unbiased and attains the CRLB. It is still possible, however, that an MVU estimator may exist. At this point we do not know how to



(a) $\hat{\theta}_1$ efficient and MVU

(b) $\hat{\theta}_1$ MVU but not efficient

Figure 3.2 Efficiency vs. minimum variance

determine whether an MVU estimator exists, and if it does, how to find it. The theory of sufficient statistics presented in Chapter 5 will allow us to answer these questions. ◇

An estimator which is unbiased and attains the CRLB, as the sample mean estimator in Example 3.3 does, is said to be *efficient* in that it efficiently uses the data. An MVU estimator may or may not be efficient. For instance, in Figure 3.2 the variances of all possible estimators (for purposes of illustration there are three unbiased estimators) are displayed. In Figure 3.2a, $\hat{\theta}_1$ is efficient in that it attains the CRLB. Therefore, it is also the MVU estimator. On the other hand, in Figure 3.2b, $\hat{\theta}_1$ does not attain the CRLB, and hence it is not efficient. However, since its variance is uniformly less than that of all other unbiased estimators, it is the MVU estimator.

The CRLB given by (3.6) may also be expressed in a slightly different form. Although (3.6) is usually more convenient for evaluation, the alternative form is sometimes useful for theoretical work. It follows from the identity (see Appendix 3A)

$$E\left[\left(\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right] \tag{3.11}$$

so that

$$\operatorname{var}(\hat{\theta}) \geq \frac{1}{E\left[\left(\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}\right)^2\right]} \tag{3.12}$$

(see Problem 3.8).

The denominator in (3.6) is referred to as the *Fisher information* $I(\theta)$ for the data $\mathbf{x}$ or

$$I(\theta) = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right]. \tag{3.13}$$

As we saw previously, when the CRLB is attained, the variance is the reciprocal of the Fisher information. Intuitively, the more information, the lower the bound. It has the essential properties of an information measure in that it is

1. nonnegative due to (3.11)
2. additive for independent observations.

The latter property leads to the result that the CRLB for N IID observations is $1/N$ times that for one observation. To verify this, note that for independent observations

$$\ln p(\mathbf{x};\theta) = \sum_{n=0}^{N-1} \ln p(x[n];\theta).$$

This results in

$$-E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right] = -\sum_{n=0}^{N-1} E\left[\frac{\partial^2 \ln p(x[n];\theta)}{\partial \theta^2}\right]$$

and finally for identically distributed observations

$$I(\theta) = Ni(\theta)$$

where

$$i(\theta) = -E\left[\frac{\partial^2 \ln p(x[n];\theta)}{\partial \theta^2}\right]$$

is the Fisher information for one sample. For nonindependent samples we might expect that the information will be less than $Ni(\theta)$, as Problem 3.9 illustrates. For completely dependent samples, as for example, $x[0] = x[1] = \cdots = x[N-1]$, we will have $I(\theta) = i(\theta)$ (see also Problem 3.9). Therefore, additional observations carry no information, and the CRLB will not decrease with increasing data record length.

3.5 General CRLB for Signals in White Gaussian Noise

Since it is common to assume white Gaussian noise, it is worthwhile to derive the CRLB for this case. Later, we will extend this to nonwhite Gaussian noise and a vector parameter as given by (3.31). Assume that a deterministic signal with an unknown parameter θ is observed in WGN as

$$x[n] = s[n;\theta] + w[n] \qquad n = 0, 1, \ldots, N-1.$$

The dependence of the signal on θ is explicitly noted. The likelihood function is

$$p(\mathbf{x};\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - s[n;\theta])^2\right\}.$$

Differentiating once produces

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n] - s[n;\theta])\frac{\partial s[n;\theta]}{\partial \theta}$$

and a second differentiation results in

$$\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left\{(x[n]-s[n;\theta])\frac{\partial^2 s[n;\theta]}{\partial\theta^2} - \left(\frac{\partial s[n;\theta]}{\partial\theta}\right)^2\right\}.$$

Taking the expected value yields

$$E\left(\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right) = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{\partial s[n;\theta]}{\partial\theta}\right)^2$$

so that finally

$$\mathrm{var}(\hat{\theta}) \geq \frac{\sigma^2}{\displaystyle\sum_{n=0}^{N-1}\left(\frac{\partial s[n;\theta]}{\partial\theta}\right)^2}. \tag{3.14}$$

The form of the bound demonstrates the importance of the signal dependence on θ. Signals that change rapidly as the unknown parameter changes result in accurate estimators. A simple application of (3.14) to Example 3.3, in which $s[n;\theta]=\theta$, produces a CRLB of σ^2/N. The reader should also verify the results of Example 3.4. As a final example we examine the problem of frequency estimation.

Example 3.5 - Sinusoidal Frequency Estimation

We assume that the signal is sinusoidal and is represented as

$$s[n;f_0] = A\cos(2\pi f_0 n + \phi) \qquad 0 < f_0 < \frac{1}{2}$$

where the amplitude and phase are known (see Example 3.14 for the case when they are unknown). From (3.14) the CRLB becomes

$$\mathrm{var}(\hat{f}_0) \geq \frac{\sigma^2}{\displaystyle A^2\sum_{n=0}^{N-1}[2\pi n \sin(2\pi f_0 n + \phi)]^2}. \tag{3.15}$$

The CRLB is plotted in Figure 3.3 versus frequency for an SNR of $A^2/\sigma^2 = 1$, a data record length of $N = 10$, and a phase of $\phi = 0$. It is interesting to note that there appear to be preferred frequencies (see also Example 3.14) for an approximation to (3.15)). Also, as $f_0 \to 0$, the CRLB goes to infinity. This is because for f_0 close to zero a slight change in frequency will not alter the signal significantly. ◇

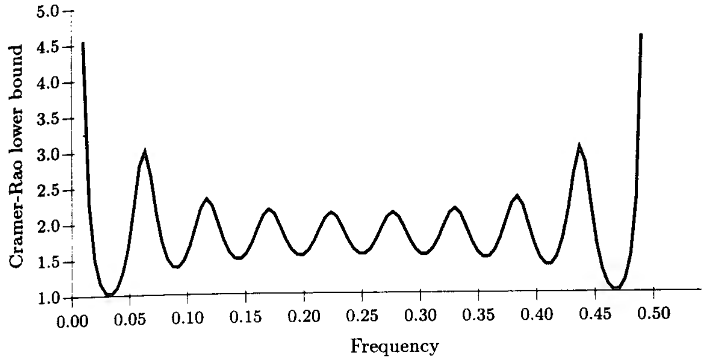


Figure 3.3 Cramer-Rao lower bound for sinusoidal frequency estimation

3.6 Transformation of Parameters

It frequently occurs in practice that the parameter we wish to estimate is a function of some more fundamental parameter. For instance, in Example 3.3 we may not be interested in the sign of A but instead may wish to estimate A^2 or the power of the signal. Knowing the CRLB for A, we can easily obtain it for A^2 or in general for any function of A. As shown in Appendix 3A, if it is desired to estimate $\alpha = g(\theta)$, then the CRLB is

$$\mathrm{var}(\hat{\alpha}) \geq \frac{\left(\dfrac{\partial g}{\partial\theta}\right)^2}{-E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right]}. \tag{3.16}$$

For the present example this becomes $\alpha = g(A) = A^2$ and

$$\boxed{\mathrm{var}(\widehat{A^2}) \geq \frac{(2A)^2}{N/\sigma^2} = \frac{4A^2\sigma^2}{N}.} \tag{3.17}$$

Note that in using (3.16) the CRLB is expressed in terms of θ.

We saw in Example 3.3 that the sample mean estimator was efficient for A. It might be supposed that $\bar{x}^2$ is efficient for A^2. To quickly dispel this notion we first show that $\bar{x}^2$ is not even an unbiased estimator. Since $\bar{x} \sim \mathcal{N}(A, \sigma^2/N)$

$$\begin{aligned} E(\bar{x}^2) &= E^2(\bar{x}) + \mathrm{var}(\bar{x}) = A^2 + \frac{\sigma^2}{N} \\ &\neq A^2. \end{aligned} \tag{3.18}$$

Hence, we immediately conclude that the *efficiency of an estimator is destroyed by a nonlinear transformation*. That it is maintained for *linear* (actually affine) transformations is easily verified. Assume that an efficient estimator for θ exists and is given by $\hat{\theta}$. It is desired to estimate $g(\theta) = a\theta + b$. As our estimator of $g(\theta)$, we choose $\widehat{g(\theta)} = g(\hat{\theta}) = a\hat{\theta} + b$. Then,

$$\begin{aligned} E(a\hat{\theta}+b) &= aE(\hat{\theta}) + b = a\theta + b \\ &= g(\theta) \end{aligned}$$

so that $\widehat{g(\theta)}$ is unbiased. The CRLB for $g(\theta)$, is from (3.16),

$$\begin{aligned} \mathrm{var}(\widehat{g(\theta)}) &\geq \frac{\left(\dfrac{\partial g}{\partial \theta}\right)^2}{I(\theta)} \\ &= \left(\frac{\partial g(\theta)}{\partial \theta}\right)^2 \mathrm{var}(\hat{\theta}) \\ &= a^2 \mathrm{var}(\hat{\theta}). \end{aligned}$$

But $\mathrm{var}(\widehat{g(\theta)}) = \mathrm{var}(a\hat{\theta}+b) = a^2\mathrm{var}(\hat{\theta})$, so that the CRLB is achieved.

Although efficiency is preserved only over linear transformations, it is *approximately* maintained over nonlinear transformations *if the data record is large enough*. This has great practical significance in that we are frequently interested in estimating functions of parameters. To see why this property holds, we return to the previous example of estimating A^2 by $\bar{x}^2$. Although $\bar{x}^2$ is biased, we note from (3.18) that $\bar{x}^2$ is *asymptotically* unbiased or unbiased as $N \to \infty$. Furthermore, since $\bar{x} \sim \mathcal{N}(A, \sigma^2/N)$, we can evaluate the variance

$$\mathrm{var}(\bar{x}^2) = E(\bar{x}^4) - E^2(\bar{x}^2)$$

by using the result that if $\xi \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\begin{aligned} E(\xi^2) &= \mu^2 + \sigma^2 \\ E(\xi^4) &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \end{aligned}$$

use variance formula
and equate it with this

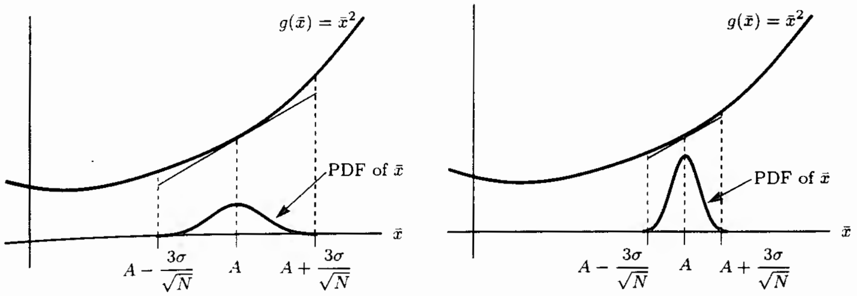
and therefore

$$\begin{aligned} \mathrm{var}(\xi^2) &= E(\xi^4) - E^2(\xi^2) \\ &= 4\mu^2\sigma^2 + 2\sigma^4. \end{aligned}$$

For our problem we have then

$$\mathrm{var}(\bar{x}^2) = \frac{4A^2\sigma^2}{N} + \frac{2\sigma^4}{N^2}. \tag{3.19}$$

Hence, as $N \to \infty$, the variance approaches $4A^2\sigma^2/N$, the last term in (3.19) converging to zero faster than the first. But this is just the CRLB as given by (3.17). Our assertion that $\bar{x}^2$ is an *asymptotically* efficient estimator of A^2 is verified. Intuitively, this situation occurs due to the *statistical linearity* of the transformation, as illustrated in Figure 3.4. As N increases, the PDF of $\bar{x}$ becomes more concentrated about the mean A. Therefore, the values of $\bar{x}$ that are observed lie in a small interval about $\bar{x} = A$ (the ±3 standard deviation interval is displayed). Over this small interval the nonlinear transformation is approximately linear. Therefore, the transformation may be replaced by a linear one since a value of $\bar{x}$ in the *nonlinear region* rarely occurs. In fact, if we linearize g about A, we have the approximation



(a) Small N **(b)** Large N

Figure 3.4 Statistical linearity of nonlinear transformations

$$g(\bar{x}) \approx g(A) + \frac{dg(A)}{dA}(\bar{x} - A).$$

It follows that, to within this approximation,

$$E[g(\bar{x})] = g(A) = A^2$$

or the estimator is unbiased (asymptotically). Also,

$$\begin{aligned} \mathrm{var}[g(\bar{x})] &= \left[\frac{dg(A)}{dA}\right]^2 \mathrm{var}(\bar{x}) \\ &= \frac{(2A)^2\sigma^2}{N} \\ &= \frac{4A^2\sigma^2}{N} \end{aligned}$$

so that the estimator achieves the CRLB (asymptotically). Therefore, it is *asymptotically efficient*. This result also yields insight into the form of the CRLB given by (3.16).

3.7 Extension to a Vector Parameter

We now extend the results of the previous sections to the case where we wish to estimate a *vector* parameter $\boldsymbol{\theta} = [\theta_1\,\theta_2 \ldots \theta_p]^T$. We will assume that the estimator $\hat{\boldsymbol{\theta}}$ is unbiased as defined in Section 2.7. The vector parameter CRLB will allow us to place a bound on the variance of each element. As derived in Appendix 3B, the CRLB is found as the $[i,i]$ element of the inverse of a matrix or

$$\mathrm{var}(\hat{\theta}_i) \geq \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii} \tag{3.20}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the $p \times p$ ***Fisher information matrix***. The latter is defined by

$$\left[\mathbf{I}(\boldsymbol{\theta})\right]_{ij} = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right] \tag{3.21}$$

for $i = 1,2,\ldots,p; j = 1,2,\ldots,p$. In evaluating (3.21) the true value of $\boldsymbol{\theta}$ is used. Note that in the scalar case ($p = 1$), $\mathbf{I}(\boldsymbol{\theta}) = I(\theta)$ and we have the scalar CRLB. Some examples follow.

Example 3.6 - DC Level in White Gaussian Noise (Revisited)

We now extend Example 3.3 to the case where in addition to A the noise variance σ^2 is also unknown. The parameter vector is $\boldsymbol{\theta} = [A\,\sigma^2]^T$, and hence $p = 2$. The 2×2 Fisher information matrix is

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2}\right] & -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial \sigma^2}\right] \\ -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \sigma^2 \partial A}\right] & -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial {\sigma^2}^2}\right] \end{bmatrix}.$$

It is clear from (3.21) that the matrix is symmetric since the order of partial differentiation may be interchanged and can also be shown to be positive definite (see Problem 3.10). The log-likelihood function is, from Example 3.3,

$$\ln p(\mathbf{x};\boldsymbol{\theta}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2.$$

The derivatives are easily found as

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A} &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) \\ \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=0}^{N-1}(x[n]-A)^2 \\ \frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2} &= -\frac{N}{\sigma^2} \\ \frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial \sigma^2} &= -\frac{1}{\sigma^4}\sum_{n=0}^{N-1}(x[n]-A) \\ \frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial {\sigma^2}^2} &= \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{n=0}^{N-1}(x[n]-A)^2. \end{aligned}$$

Upon taking the negative expectations, the Fisher information matrix becomes

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \dfrac{N}{\sigma^2} & 0 \\ 0 & \dfrac{N}{2\sigma^4} \end{bmatrix}.$$

Although not true in general, for this example the Fisher information matrix is diagonal and hence easily inverted to yield

$$\begin{aligned} \mathrm{var}(\hat{A}) &\geq \frac{\sigma^2}{N} \\ \mathrm{var}(\hat{\sigma^2}) &\geq \frac{2\sigma^4}{N}. \end{aligned}$$

Note that the CRLB for $\hat{A}$ is the same as for the case when σ^2 is known due to the diagonal nature of the matrix. Again this is not true in general, as the next example illustrates. ◇

Example 3.7 - Line Fitting

Consider the problem of line fitting or given the observations

$$x[n] = A + Bn + w[n] \qquad n = 0,1,\ldots,N-1$$

where $w[n]$ is WGN, determine the CRLB for the slope B and the intercept A. The parameter vector in this case is $\boldsymbol{\theta} = [A\,B]^T$. We need to first compute the 2×2 Fisher information matrix,

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2}\right] & -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial B}\right] \\ -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial B \partial A}\right] & -E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial B^2}\right] \end{bmatrix}.$$

The likelihood function is

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A-Bn)^2\right\}$$

from which the derivatives follow as

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A} &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A-Bn) \\ \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial B} &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A-Bn)n \end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2} &= -\frac{N}{\sigma^2} \\
\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial B} &= -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} n \\
\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial B^2} &= -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} n^2.
\end{aligned}$$

Since the second-order derivatives do not depend on $\mathbf{x}$, we have immediately that

$$\begin{aligned}
\mathbf{I}(\boldsymbol{\theta}) &= \frac{1}{\sigma^2}\begin{bmatrix} N & \sum_{n=0}^{N-1} n \\ \sum_{n=0}^{N-1} n & \sum_{n=0}^{N-1} n^2 \end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix} N & \frac{N(N-1)}{2} \\ \frac{N(N-1)}{2} & \frac{N(N-1)(2N-1)}{6} \end{bmatrix}
\end{aligned}$$

where we have used the identities

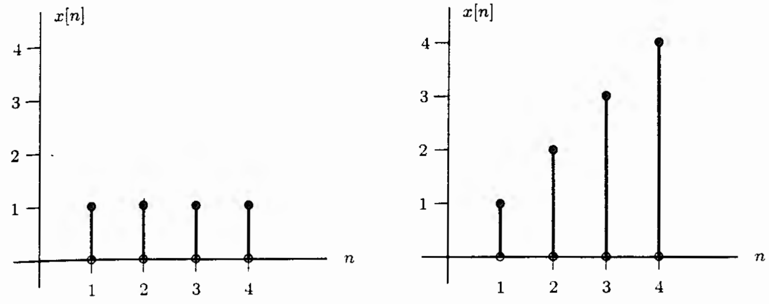
$$\begin{aligned}
\sum_{n=0}^{N-1} n &= \frac{N(N-1)}{2} \\
\sum_{n=0}^{N-1} n^2 &= \frac{N(N-1)(2N-1)}{6}.
\end{aligned} \tag{3.22}$$

Inverting the matrix yields

$$\mathbf{I}^{-1}(\boldsymbol{\theta}) = \sigma^2 \begin{bmatrix} \frac{2(2N-1)}{N(N+1)} & -\frac{6}{N(N+1)} \\ -\frac{6}{N(N+1)} & \frac{12}{N(N^2-1)} \end{bmatrix}.$$

It follows from (3.20) that the CRLB is

$$\begin{aligned}
\operatorname{var}(\hat{A}) &\geq \frac{2(2N-1)\sigma^2}{N(N+1)} \\
\operatorname{var}(\hat{B}) &\geq \frac{12\sigma^2}{N(N^2-1)}.
\end{aligned}$$



(a) $A = 0,\ B = 0$ to $A = 1,\ B = 0$ (b) $A = 0,\ B = 0$ to $A = 0,\ B = 1$

Figure 3.5 Sensitivity of observations to parameter changes—no noise

Some interesting observations follow from examination of the CRLB. Note first that the CRLB for $\hat{A}$ has increased over that obtained when B is known, for in the latter case we have

$$\operatorname{var}(\hat{A}) \geq -\frac{1}{E\left[\frac{\partial^2 \ln p(\mathbf{x};A)}{\partial A^2}\right]} = \frac{\sigma^2}{N}$$

and for $N \geq 2$, $2(2N-1)/(N+1) > 1$. This is a quite general result that asserts that ***the CRLB always increases as we estimate more parameters*** (see Problems 3.11 and 3.12). A second point is that

$$\frac{\operatorname{CRLB}(\hat{A})}{\operatorname{CRLB}(\hat{B})} = \frac{(2N-1)(N-1)}{6} > 1$$

for $N \geq 3$. Hence, B is easier to estimate, its CRLB decreasing as $1/N^3$ as opposed to the $1/N$ dependence for the CRLB of A. These differing dependences indicate that $x[n]$ is *more sensitive* to changes in B than to changes in A. A simple calculation reveals

$$\begin{aligned}
\Delta x[n] &\approx \frac{\partial x[n]}{\partial A}\Delta A = \Delta A \\
\Delta x[n] &\approx \frac{\partial x[n]}{\partial B}\Delta B = n\Delta B.
\end{aligned}$$

Changes in B are magnified by n, as illustrated in Figure 3.5. This effect is reminiscent of (3.14), and indeed a similar type of relationship is obtained in the vector parameter case (see (3.33)). See Problem 3.13 for a generalization of this example. ◇

As an alternative means of computing the CRLB we can use the identity

$$E\left[\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_i}\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_j}\right] = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right] \tag{3.23}$$

as shown in Appendix 3B. The form given on the right-hand side is usually easier to evaluate, however.

We now formally state the CRLB theorem for a vector parameter. Included in the theorem are conditions for equality. The bound is stated in terms of the covariance matrix of $\hat{\boldsymbol{\theta}}$, denoted by $\mathbf{C}_{\hat{\theta}}$, from which (3.20) follows.

Theorem 3.2 (Cramer-Rao Lower Bound - Vector Parameter) *It is assumed that the PDF $p(\mathbf{x};\boldsymbol{\theta})$ satisfies the "regularity" conditions*

$$E\left[\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0} \qquad \textit{for all } \boldsymbol{\theta}$$

where the expectation is taken with respect to $p(\mathbf{x};\boldsymbol{\theta})$. Then, the covariance matrix of any unbiased estimator $\hat{\boldsymbol{\theta}}$ satisfies

$$\mathbf{C}_{\hat{\theta}} - \mathbf{I}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0} \tag{3.24}$$

where $\geq \mathbf{0}$ is interpreted as meaning that the matrix is positive semidefinite. The Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ is given as

$$[\mathbf{I}(\boldsymbol{\theta})]_{ij} = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right]$$

where the derivatives are evaluated at the true value of $\boldsymbol{\theta}$ and the expectation is taken with respect to $p(\mathbf{x};\boldsymbol{\theta})$. Furthermore, an unbiased estimator may be found that attains the bound in that $\mathbf{C}_{\hat{\theta}} = \mathbf{I}^{-1}(\boldsymbol{\theta})$ if and only if

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta}) \tag{3.25}$$

for some p-dimensional function $\mathbf{g}$ and some $p \times p$ matrix $\mathbf{I}$. That estimator, which is the MVU estimator, is $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$, and its covariance matrix is $\mathbf{I}^{-1}(\boldsymbol{\theta})$.

The proof is given in Appendix 3B. That (3.20) follows from (3.24) is shown by noting that for a positive semidefinite matrix the diagonal elements are nonnegative. Hence,

$$\left[\mathbf{C}_{\hat{\theta}} - \mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii} \geq 0$$

and therefore

$$\operatorname{var}(\hat{\theta}_i) = [\mathbf{C}_{\hat{\theta}}]_{ii} \geq \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}. \tag{3.26}$$

Additionally, when equality holds or $\mathbf{C}_{\hat{\theta}} = \mathbf{I}^{-1}(\boldsymbol{\theta})$, then (3.26) holds with equality also. The conditions for the CRLB to be attained are of particular interest since then $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$ is efficient and hence is the MVU estimator. An example of equality occurs in Example 3.7. There we found that

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \dfrac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A} \\ \dfrac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial B} \end{bmatrix} \tag{3.27}$$

$$= \begin{bmatrix} \dfrac{1}{\sigma^2}\displaystyle\sum_{n=0}^{N-1}(x[n] - A - Bn) \\ \dfrac{1}{\sigma^2}\displaystyle\sum_{n=0}^{N-1}(x[n] - A - Bn)n \end{bmatrix}. \tag{3.28}$$

Although not obvious, this may be rewritten as

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \dfrac{N}{\sigma^2} & \dfrac{N(N-1)}{2\sigma^2} \\ \dfrac{N(N-1)}{2\sigma^2} & \dfrac{N(N-1)(2N-1)}{6\sigma^2} \end{bmatrix} \begin{bmatrix} \hat{A} - A \\ \hat{B} - B \end{bmatrix} \tag{3.29}$$

where

$$\begin{aligned} \hat{A} &= \frac{2(2N-1)}{N(N+1)}\sum_{n=0}^{N-1} x[n] - \frac{6}{N(N+1)}\sum_{n=0}^{N-1} nx[n] \\ \hat{B} &= -\frac{6}{N(N+1)}\sum_{n=0}^{N-1} x[n] + \frac{12}{N(N^2-1)}\sum_{n=0}^{N-1} nx[n]. \end{aligned}$$

Hence, the conditions for equality are satisfied and $[\hat{A}\ \hat{B}]^T$ is an efficient and therefore MVU estimator. Furthermore, the matrix in (3.29) is the inverse of the covariance matrix.

If the equality conditions hold, the reader may ask whether we can be assured that $\hat{\boldsymbol{\theta}}$ is unbiased. Because the regularity conditions

$$E\left[\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}$$

are always assumed to hold, we can apply them to (3.25). This then yields $E[\mathbf{g}(\mathbf{x})] = E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$.

In finding MVU estimators for a vector parameter the CRLB theorem provides a powerful tool. In particular, it allows us to find the MVU estimator for an important class of data models. This class is the ***linear model*** and is described in detail in Chapter 4. The line fitting example just discussed is a special case. Suffice it to say that if we can model our data in the linear model form, then the MVU estimator and its performance are easily found.

3.8 Vector Parameter CRLB for Transformations

The discussion in Section 3.6 extends readily to the vector case. Assume that it is desired to estimate $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$ for $\mathbf{g}$, an r-dimensional function. Then, as shown in Appendix 3B

$$\mathbf{C}_{\hat{\alpha}} - \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\mathbf{I}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{g}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} \geq \mathbf{0} \tag{3.30}$$

where, as before, $\geq \mathbf{0}$ is to be interpreted as positive semidefinite. In (3.30), $\partial\mathbf{g}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is the $r \times p$ **Jacobian matrix** defined as

$$\frac{\partial\mathbf{g}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial g_1(\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial g_1(\boldsymbol{\theta})}{\partial\theta_2} & \cdots & \frac{\partial g_1(\boldsymbol{\theta})}{\partial\theta_p} \\ \frac{\partial g_2(\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial g_2(\boldsymbol{\theta})}{\partial\theta_2} & \cdots & \frac{\partial g_2(\boldsymbol{\theta})}{\partial\theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_r(\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial g_r(\boldsymbol{\theta})}{\partial\theta_2} & \cdots & \frac{\partial g_r(\boldsymbol{\theta})}{\partial\theta_p} \end{bmatrix}.$$

Example 3.8 - CRLB for Signal-to-Noise Ratio

Consider a DC level in WGN with A and σ^2 unknown. We wish to estimate

$$\alpha = \frac{A^2}{\sigma^2}$$

which can be considered to be the SNR for a single sample. Here $\boldsymbol{\theta} = [A\,\sigma^2]^T$ and $g(\boldsymbol{\theta}) = \theta_1^2/\theta_2 = A^2/\sigma^2$. Then, as shown in Example 3.6,

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}.$$

The Jacobian is

$$\begin{aligned} \frac{\partial g(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} &= \begin{bmatrix} \frac{\partial g(\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial g(\boldsymbol{\theta})}{\partial\theta_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial g(\boldsymbol{\theta})}{\partial A} & \frac{\partial g(\boldsymbol{\theta})}{\partial\sigma^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{2A}{\sigma^2} & -\frac{A^2}{\sigma^4} \end{bmatrix} \end{aligned}$$

so that

$$\begin{aligned} \frac{\partial\mathbf{g}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{I}^{-1}(\boldsymbol{\theta})\frac{\partial\mathbf{g}(\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}} &= \begin{bmatrix} \frac{2A}{\sigma^2} & -\frac{A^2}{\sigma^4} \end{bmatrix} \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix} \begin{bmatrix} \frac{2A}{\sigma^2} \\ -\frac{A^2}{\sigma^4} \end{bmatrix} \\ &= \frac{4A^2}{N\sigma^2} + \frac{2A^4}{N\sigma^4} \\ &= \frac{4\alpha + 2\alpha^2}{N}. \end{aligned}$$

Finally, since α is a scalar

$$\operatorname{var}(\hat{\alpha}) \geq \frac{4\alpha + 2\alpha^2}{N}.$$

◇

As discussed in Section 3.6, efficiency is maintained over linear transformations

$$\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta} + \mathbf{b}$$

where $\mathbf{A}$ is an $r \times p$ matrix and $\mathbf{b}$ is an $r \times 1$ vector. If $\hat{\boldsymbol{\alpha}} = \mathbf{A}\hat{\boldsymbol{\theta}} + \mathbf{b}$, and $\hat{\boldsymbol{\theta}}$ is efficient or $\mathbf{C}_{\hat{\theta}} = \mathbf{I}^{-1}(\boldsymbol{\theta})$, then

$$E(\hat{\boldsymbol{\alpha}}) = \mathbf{A}\boldsymbol{\theta} + \mathbf{b} = \boldsymbol{\alpha}$$

so that $\hat{\boldsymbol{\alpha}}$ is unbiased and

$$\begin{aligned} \mathbf{C}_{\hat{\alpha}} &= \mathbf{A}\mathbf{C}_{\hat{\theta}}\mathbf{A}^T = \mathbf{A}\mathbf{I}^{-1}(\boldsymbol{\theta})\mathbf{A}^T \\ &= \frac{\partial\mathbf{g}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{I}^{-1}(\boldsymbol{\theta})\frac{\partial\mathbf{g}(\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}}, \end{aligned}$$

the latter being the CRLB. For nonlinear transformations efficiency is maintained only as $N \to \infty$. (This assumes that the PDF of $\hat{\boldsymbol{\theta}}$ becomes concentrated about the true value of $\boldsymbol{\theta}$ as $N \to \infty$ or that $\hat{\boldsymbol{\theta}}$ is consistent.) Again this is due to the statistical linearity of $\mathbf{g}(\boldsymbol{\theta})$ about the true value of $\boldsymbol{\theta}$.

3.9 CRLB for the General Gaussian Case

It is quite convenient at times to have a general expression for the CRLB. In the case of Gaussian observations we can derive the CRLB that generalizes (3.14). Assume that

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{C}(\boldsymbol{\theta}))$$

so that both the mean and covariance may depend on $\boldsymbol{\theta}$. Then, as shown in Appendix 3C, the Fisher information matrix is given by

$$\boxed{\begin{aligned} [\mathbf{I}(\boldsymbol{\theta})]_{ij} &= \left[\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_i}\right]^T \mathbf{C}^{-1}(\boldsymbol{\theta}) \left[\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_j}\right] \\ &+ \frac{1}{2}\operatorname{tr}\left[\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial\mathbf{C}(\boldsymbol{\theta})}{\partial\theta_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial\mathbf{C}(\boldsymbol{\theta})}{\partial\theta_j}\right] \end{aligned}} \tag{3.31}$$

where

$$\boxed{\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_i} = \begin{bmatrix} \frac{\partial[\boldsymbol{\mu}(\boldsymbol{\theta})]_1}{\partial\theta_i} \\ \frac{\partial[\boldsymbol{\mu}(\boldsymbol{\theta})]_2}{\partial\theta_i} \\ \vdots \\ \frac{\partial[\boldsymbol{\mu}(\boldsymbol{\theta})]_N}{\partial\theta_i} \end{bmatrix}}$$

and

$$\boxed{\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_i} = \begin{bmatrix} \frac{\partial[\mathbf{C}(\boldsymbol{\theta})]_{11}}{\partial \theta_i} & \frac{\partial[\mathbf{C}\ \theta \ [illegible]}{[illegible]} & \cdots & \frac{\partial[\mathbf{C}(\boldsymbol{\theta})]_{1N}}{\partial \theta_i} \\ \frac{\partial[\mathbf{C}(\boldsymbol{\theta})]_{21}}{\partial \theta_i} & \frac{\partial[\mathbf{C}\ \theta \ [illegible]}{\partial [illegible]} & \cdots & \frac{\partial[\mathbf{C}(\boldsymbol{\theta})]_{2N}}{\partial \theta_i} \\ \vdots & & & \vdots \\ \frac{\partial[\mathbf{C}(\boldsymbol{\theta})]_{N1}}{\partial \theta_i} & \frac{\partial[\mathbf{C}\ \theta \ [illegible]}{\partial [illegible]} & \cdots & \frac{\partial[\mathbf{C}(\boldsymbol{\theta})]_{NN}}{\partial \theta_i} \end{bmatrix}}.$$

For the scalar parameter case in which

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}(\theta \ [illegible]$$

this reduces to

$$\begin{aligned} I(\theta) &= \left[\frac{\partial \boldsymbol{\mu}(\theta)}{\partial \theta}\right]^T \mathbf{C}^{[illegible]} \left[\frac{[illegible]\boldsymbol{\mu}(\theta)}{[illegible]\theta}\right] \\ &\quad + \frac{1}{2}\mathrm{tr}\left[\left(\mathbf{C}^{[illegible]} \frac{[illegible]\mathbf{C}(\theta)}{[illegible]\theta}\right)^2\right] \end{aligned} \qquad (3.32)$$

which generalizes (3.14). We now illustrate th[illegible] [illegible]ation with some examples.

Example 3.9 - Parameters of a Signal in [illegible]e Gaussian Noise

Assume that we wish to estimate a scalar signa[illegible] [illegible]eter θ for the data set

$$x[n] = s[n;\theta] + w[n] \qquad [illegible] = [illegible] \ldots, N-1$$

where $w[n]$ is WGN. The covariance matrix is $\mathbf{C} = [illegible]\mathbf{I}$ and does not depend on θ. The second term in (3.32) is therefore zero. The fi[illegible] [illegible] [illegible]elds

$$\begin{aligned} I(\theta) &= \frac{1}{\sigma^2}\left[\frac{\partial \boldsymbol{\mu}(\theta [illegible]}{\partial \theta} [illegible] \frac{[illegible]\boldsymbol{\mu}(\theta)}{[illegible]}\right] \\ &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{[illegible]}{[illegible]}\right. [illegible] \\ &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{\partial s[illegible]}{[illegible]}\right. [illegible] \end{aligned}$$

which agrees with (3.14). ◇

Generalizing to a vector signal parameter estima[illegible] [illegible] [illegible]e presence of WGN, we have from (3.31)

$$[\mathbf{I}(\boldsymbol{\theta})]_{ij} = \left[\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i}\right]^T [illegible] \left[\frac{[illegible]\boldsymbol{\mu}(\boldsymbol{\theta})}{[illegible]}\right]$$

which yields

$$[\mathbf{I}(\boldsymbol{\theta})]_{ij} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\frac{\partial s[n;\boldsymbol{\theta}]}{\partial \theta_i}\frac{\partial s[n;\boldsymbol{\theta}]}{\partial \theta_j} \qquad (3.33)$$

as the elements of the Fisher information matrix.

Example 3.10 - Parameter of Noise

Assume that we observe

$$x[n] = w[n] \qquad n = 0, 1, \ldots, N-1$$

where $w[n]$ is WGN with unknown variance $\theta = \sigma^2$. Then, according to (3.32), since $\mathbf{C}(\sigma^2) = \sigma^2\mathbf{I}$, we have

$$\begin{aligned} I(\sigma^2) &= \frac{1}{2}\mathrm{tr}\left[\left(\mathbf{C}^{-1}(\sigma^2)\frac{\partial \mathbf{C}(\sigma^2)}{\partial \sigma^2}\right)^2\right] \\ &= \frac{1}{2}\mathrm{tr}\left[\left(\left(\frac{1}{\sigma^2}\right)(\mathbf{I})\right)^2\right] \\ &= \frac{1}{2}\mathrm{tr}\left[\frac{1}{\sigma^4}\mathbf{I}\right] \\ &= \frac{N}{2\sigma^4} \end{aligned}$$

which agrees with the results in Example 3.6. A slightly more complicated example follows. ◇

Example 3.11 - Random DC Level in WGN

Consider the data

$$x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $w[n]$ is WGN and A, the DC level, is a Gaussian *random variable* with zero mean and variance σ_A^2. Also, A is independent of $w[n]$. The power of the signal or variance σ_A^2 is the unknown parameter. Then, $\mathbf{x} = [x[0]\, x[1] \ldots x[N-1]]^T$ is Gaussian with zero mean and an $N \times N$ covariance matrix whose $[i,j]$ element is

$$\begin{aligned} [\mathbf{C}(\sigma_A^2)]_{ij} &= E\left[x[i-1]x[j-1]\right] \\ &= E\left[(A + w[i-1])(A + w[j-1])\right] \\ &= \sigma_A^2 + \sigma^2\delta_{ij}. \end{aligned}$$

Therefore,

$$\mathbf{C}(\sigma_A^2) = \sigma_A^2\mathbf{1}\mathbf{1}^T + \sigma^2\mathbf{I}$$

Chapter 4

Linear Models

4.1 Introduction

The determination of the MVU estimator is *in general* a difficult task. It is fortunate, however, that a large number of signal processing estimation problems can be represented by a data model that allows us to easily determine this estimator. This class of models is **the *linear model*.** Not only is the MVU estimator immediately evident once the linear model has been identified, but in addition, the statistical performance follows naturally. The key, then, to finding the optimal estimator is in structuring the problem in the linear model form to take advantage of its unique properties.

4.2 Summary

The linear model is defined by (4.8). When this data model can be assumed, the MVU (and also efficient) estimator is given by (4.9), and the covariance matrix by (4.10). A more general model, termed the general linear model, allows the noise to have an arbitrary covariance matrix, as opposed to $\sigma^2\mathbf{I}$ for the linear model. The MVU (and also efficient) estimator for this model is given by (4.25), and its corresponding covariance matrix by (4.26). A final extension allows for known signal components in the data to yield the MVU (and also efficient) estimator of (4.31). The covariance matrix is the same as for the general linear model.

4.3 Definition and Properties

The linear model has already been encountered in the line fitting problem discussed in Example 3.7. Recall that the problem was to fit a straight line through noise corrupted data. As our model of the data we chose

$$x[n] = A + Bn + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $w[n]$ is WGN and the slope B and intercept A were to be estimated. In matrix notation the model is written more compactly as

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{4.1}$$

where

$$\begin{aligned} \mathbf{x} &= [x[0]\,x[1]\dots x[N-1]]^T \\ \mathbf{w} &= [w[0]\,w[1]\dots w[N-1]]^T \\ \boldsymbol{\theta} &= [A\,B]^T \end{aligned}$$

and

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & N-1 \end{bmatrix}.$$

The matrix $\mathbf{H}$ is a *known* matrix of dimension $N \times 2$ and is referred to as the ***observation matrix***. The data $\mathbf{x}$ are observed after $\boldsymbol{\theta}$ is operated upon by $\mathbf{H}$. Note also that the noise vector has the statistical characterization $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The data model in (4.1) is termed the ***linear model***. In defining the linear model we assume that the noise vector is Gaussian, although other authors use the term more generally for any noise PDF [Graybill 1976].

As discussed in Chapter 3, it is sometimes possible to determine the MVU estimator if the equality constraints of the CRLB theorem are satisfied. From Theorem 3.2, $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$ will be the MVU estimator if

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta}) \tag{4.2}$$

for some function $\mathbf{g}$. Furthermore, the covariance matrix of $\hat{\boldsymbol{\theta}}$ will be $\mathbf{I}^{-1}(\boldsymbol{\theta})$. To determine if this condition is satisfied for the linear model of (4.1), we have

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}}\left[-\ln(2\pi\sigma^2)^{\frac{N}{2}} - \frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})\right] \\ &= -\frac{1}{2\sigma^2}\frac{\partial}{\partial \boldsymbol{\theta}}\left[\mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}\right]. \end{aligned}$$

Using the identities

$$\begin{aligned} \frac{\partial \mathbf{b}^T\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} &= \mathbf{b} \\ \frac{\partial \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} &= 2\mathbf{A}\boldsymbol{\theta} \end{aligned} \tag{4.3}$$

for $\mathbf{A}$ a symmetric matrix, we have

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{\sigma^2}[\mathbf{H}^T\mathbf{x} - \mathbf{H}^T\mathbf{H}\boldsymbol{\theta}].$$

Assuming that $\mathbf{H}^T\mathbf{H}$ is invertible

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}[(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} - \boldsymbol{\theta}] \tag{4.4}$$

which is exactly in the form of (4.2) with

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \tag{4.5}$$

$$\mathbf{I}(\boldsymbol{\theta}) = \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}. \tag{4.6}$$

Hence, the MVU estimator of $\boldsymbol{\theta}$ is given by (4.5), and its covariance matrix is

$$\mathbf{C}_{\hat{\theta}} = \mathbf{I}^{-1}(\boldsymbol{\theta}) = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}. \tag{4.7}$$

Additionally, the MVU estimator for the linear model is efficient in that it attains the CRLB. The reader may now verify the result of (3.29) by substituting $\mathbf{H}$ for the line fitting problem into (4.4). The only detail that requires closer scrutiny is the invertibility of $\mathbf{H}^T\mathbf{H}$. For the line fitting example a direct calculation will verify that the inverse exists (compute the determinant of the matrix given in (3.29)). Alternatively, this follows from the linear independence of the columns of $\mathbf{H}$ (see Problem 4.2). If the columns of $\mathbf{H}$ are not linearly independent, as for example,

$$\mathbf{H} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

and $\mathbf{x} = [2\,2\dots 2]^T$ so that $\mathbf{x}$ lies in the range space of $\mathbf{H}$, then *even in the absence of noise the model parameters will not be identifiable.* For then

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta}$$

and for this choice of $\mathbf{H}$ we will have for $x[n]$

$$2 = A + B \qquad n = 0, 1, \dots, N-1.$$

As illustrated in Figure 4.1 it is clear that an infinite number of choices can be made for A and B that will result in the same observations or given a noiseless $\mathbf{x}$, $\boldsymbol{\theta}$ is not unique. The situation can hardly hope to improve when the observations are corrupted by noise. Although rarely occurring in practice, this degeneracy sometimes *nearly occurs* when $\mathbf{H}^T\mathbf{H}$ is ill-conditioned (see Problem 4.3).

The previous discussion, although illustrated by the line fitting example, is completely general, as summarized by the following theorem.

Theorem 4.1 (Minimum Variance Unbiased Estimator for the Linear Model) *If the data observed can be modeled as*

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{4.8}$$

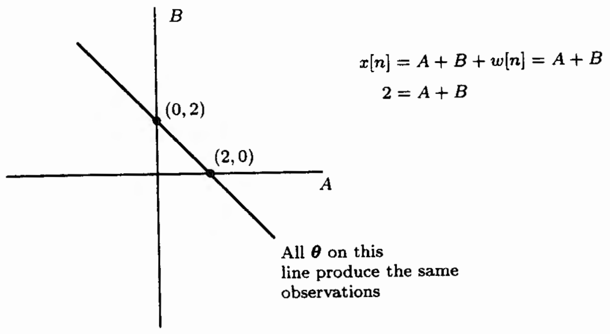

B
(0, 2)
(2, 0)
A
$x[n] = A + B + w[n] = A + B$
$2 = A + B$
All $\boldsymbol{\theta}$ on this line produce the same observations

Figure 4.1 Nonidentifiability of linear model parameters

where $\mathbf{x}$ *is an* $N \times 1$ *vector of observations,* $\mathbf{H}$ *is a known* $N \times p$ *observation matrix (with* $N > p$*) and rank* p*,* $\boldsymbol{\theta}$ *is a* $p \times 1$ *vector of parameters to be estimated, and* $\mathbf{w}$ *is an* $N \times 1$ *noise vector with PDF* $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$*, then the MVU estimator is*

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \tag{4.9}$$

and the covariance matrix of $\hat{\boldsymbol{\theta}}$ *is*

$$\mathbf{C}_{\hat{\theta}} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}. \tag{4.10}$$

For the linear model the MVU estimator is efficient in that it attains the CRLB.

That $\hat{\boldsymbol{\theta}}$ is unbiased easily follows by substituting (4.8) into (4.9). Also, the statistical performance of $\hat{\boldsymbol{\theta}}$ is *completely specified* (not just the mean and covariance) because $\hat{\boldsymbol{\theta}}$ is a *linear transformation* of a Gaussian vector $\mathbf{x}$ and hence

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}). \tag{4.11}$$

The Gaussian nature of the MVU estimator for the linear model allows us to determine the exact statistical performance if desired (see Problem 4.4). In the next section we present some examples illustrating the use of the linear model.

4.4 Linear Model Examples

We have already seen how the problem of line fitting is easily handled once we recognize it as a linear model. A simple extension is to the problem of fitting a curve to experimental data.

Example 4.1 - Curve Fitting

In many experimental situations we seek to determine an empirical relationship between a pair of variables. For instance, in Figure 4.2 we present the results of a experiment

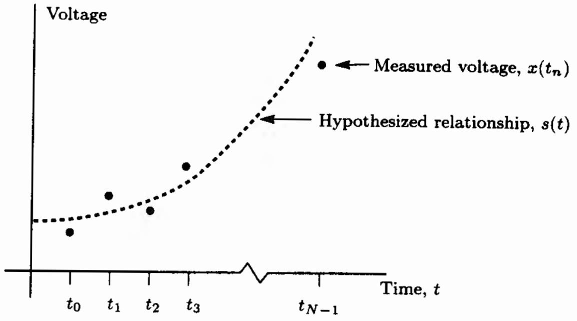


Figure 4.2 Experimental data

in which voltage measurements are taken at the time instants $t = t_0, t_1, \ldots, t_{N-1}$. By plotting the measurements it appears that the underlying voltage may be a quadratic function of time. That the points do not lie exactly on a curve is attributed to experimental error or noise. Hence, a reasonable model for the data is

$$x(t_n) = \theta_1 + \theta_2 t_n + \theta_3 t_n^2 + w(t_n) \qquad n = 0, 1, \ldots, N-1.$$

To avail ourselves of the utility of the linear model we assume that $w(t_n)$ are IID Gaussian random variables with zero mean and variance σ^2 or that they are WGN samples. Then, we have the usual linear model form

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

where

$$\begin{aligned} \mathbf{x} &= [x(t_0)\, x(t_1) \ldots x(t_{N-1})]^T \\ \boldsymbol{\theta} &= [\theta_1\, \theta_2\, \theta_3]^T \end{aligned}$$

and

$$\mathbf{H} = \begin{bmatrix} 1 & t_0 & t_0^2 \\ 1 & t_1 & t_1^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{N-1} & t_{N-1}^2 \end{bmatrix}.$$

In general, if we seek to fit a $(p-1)$st-order polynomial to experimental data we will have

$$x(t_n) = \theta_1 + \theta_2 t_n + \cdots + \theta_p t_n^{p-1} + w(t_n) \qquad n = 0, 1, \ldots, N-1.$$

The MVU estimator for $\boldsymbol{\theta} = [\theta_1\, \theta_2 \ldots \theta_p]^T$ follows from (4.9) as

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

where

$$\begin{aligned}\mathbf{x} &= [x(t_0)\, x(t_1) \dots x(t_{N-1})]^T \\ \mathbf{H} &= \begin{bmatrix} 1 & t_0 & \dots & t_0^{p-1} \\ 1 & t_1 & \dots & t_1^{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{N-1} & \dots & t_{N-1}^{p-1} \end{bmatrix} \quad (N \times p).\end{aligned}$$

The observation matrix for this example has the special form of a Vandermonde matrix. Note that the resultant curve fit is

$$\hat{s}(t) = \sum_{i=1}^{p} \hat{\theta}_i t^{i-1}$$

where $s(t)$ denotes the underlying curve or signal. ◇

Example 4.2 - Fourier Analysis

Many signals exhibit cyclical behavior. It is common practice to determine the presence of strong cyclical components by employing a Fourier analysis. Large Fourier coefficients are indicative of strong components. In this example we show that a Fourier analysis is really just an estimation of the linear model parameters. Consider a data model consisting of sinusoids in white Gaussian noise:

$$x[n] = \sum_{k=1}^{M} a_k \cos\left(\frac{2\pi kn}{N}\right) + \sum_{k=1}^{M} b_k \sin\left(\frac{2\pi kn}{N}\right) + w[n] \qquad n = 0, 1, \dots, N-1 \quad (4.12)$$

where $w[n]$ is WGN. The frequencies are assumed to be harmonically related or multiples of the fundamental $f_1 = 1/N$ as $f_k = k/N$. The amplitudes a_k, b_k of the cosines and sines are to be estimated. To reformulate the problem in terms of the linear model we let

$$\boldsymbol{\theta} = [a_1\, a_2 \dots a_M\, b_1\, b_2 \dots b_M]^T$$

and

$$\mathbf{H} = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ \cos\left(\frac{2\pi}{N}\right) & \dots & \cos\left(\frac{2\pi M}{N}\right) & \sin\left(\frac{2\pi}{N}\right) & \dots & \sin\left(\frac{2\pi M}{N}\right) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos\left[\frac{2\pi(N-1)}{N}\right] & \dots & \cos\left[\frac{2\pi M(N-1)}{N}\right] & \sin\left[\frac{2\pi(N-1)}{N}\right] & \dots & \sin\left[\frac{2\pi M(N-1)}{N}\right] \end{bmatrix}.$$

Note that $\mathbf{H}$ has dimension $N \times 2M$, where $p = 2M$. Hence, for $\mathbf{H}$ to satisfy $N > p$ we require $M < N/2$. In determining the MVU estimator we can simplify the computations by noting that the columns of $\mathbf{H}$ are orthogonal. Let $\mathbf{H}$ be represented in column form as

$$\mathbf{H} = [\mathbf{h}_1\, \mathbf{h}_2 \dots \mathbf{h}_{2M}]$$

where $\mathbf{h}_i$ denotes the ith column of $\mathbf{H}$. Then, it follows that

$$\mathbf{h}_i^T \mathbf{h}_j = 0 \qquad \text{for } i \neq j.$$

This property is quite useful in that

$$\begin{aligned}\mathbf{H}^T\mathbf{H} &= \begin{bmatrix} \mathbf{h}_1^T \\ \vdots \\ \mathbf{h}_{2M}^T \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 & \dots & \mathbf{h}_{2M} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{h}_1^T\mathbf{h}_1 & \mathbf{h}_1^T\mathbf{h}_2 & \dots & \mathbf{h}_1^T\mathbf{h}_{2M} \\ \mathbf{h}_2^T\mathbf{h}_1 & \mathbf{h}_2^T\mathbf{h}_2 & \dots & \mathbf{h}_2^T\mathbf{h}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{2M}^T\mathbf{h}_1 & \mathbf{h}_{2M}^T\mathbf{h}_2 & \dots & \mathbf{h}_{2M}^T\mathbf{h}_{2M} \end{bmatrix}\end{aligned}$$

becomes a diagonal matrix which is easily inverted. The orthogonality of the columns results from the discrete Fourier transform (DFT) relationships for $i, j = 1, 2, \dots, M < N/2$:

$$\boxed{\begin{aligned} \sum_{n=0}^{N-1} \cos\left(\frac{2\pi in}{N}\right) \cos\left(\frac{2\pi jn}{N}\right) &= \frac{N}{2}\delta_{ij} \\ \sum_{n=0}^{N-1} \sin\left(\frac{2\pi in}{N}\right) \sin\left(\frac{2\pi jn}{N}\right) &= \frac{N}{2}\delta_{ij} \\ \sum_{n=0}^{N-1} \cos\left(\frac{2\pi in}{N}\right) \sin\left(\frac{2\pi jn}{N}\right) &= 0 \quad \text{for all } i, j. \end{aligned}} \qquad (4.13)$$

An outline of the orthogonality proof is given in Problem 4.5. Using this property, we have

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} \frac{N}{2} & 0 & \dots & 0 \\ 0 & \frac{N}{2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{N}{2} \end{bmatrix} = \frac{N}{2}\mathbf{I}$$

so that the MVU estimator of the amplitudes is

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}\\ &= \frac{2}{N}\mathbf{H}^T\mathbf{x} = \frac{2}{N}\begin{bmatrix}\mathbf{h}_1^T\\ \vdots\\ \mathbf{h}_{2M}^T\end{bmatrix}\mathbf{x}\\ &= \begin{bmatrix}\frac{2}{N}\mathbf{h}_1^T\mathbf{x}\\ \vdots\\ \frac{2}{N}\mathbf{h}_{2M}^T\mathbf{x}\end{bmatrix}\end{aligned}$$

or finally,

$$\boxed{\begin{aligned}\hat{a}_k &= \frac{2}{N}\sum_{n=0}^{N-1} x[n]\cos\left(\frac{2\pi kn}{N}\right)\\ \hat{b}_k &= \frac{2}{N}\sum_{n=0}^{N-1} x[n]\sin\left(\frac{2\pi kn}{N}\right).\end{aligned}} \qquad (4.14)$$

These are recognized as the discrete Fourier transform coefficients. From the properties of the linear model we can immediately conclude that the means are

$$\begin{aligned}E(\hat{a}_k) &= a_k\\ E(\hat{b}_k) &= b_k\end{aligned} \qquad (4.15)$$

and the covariance matrix is

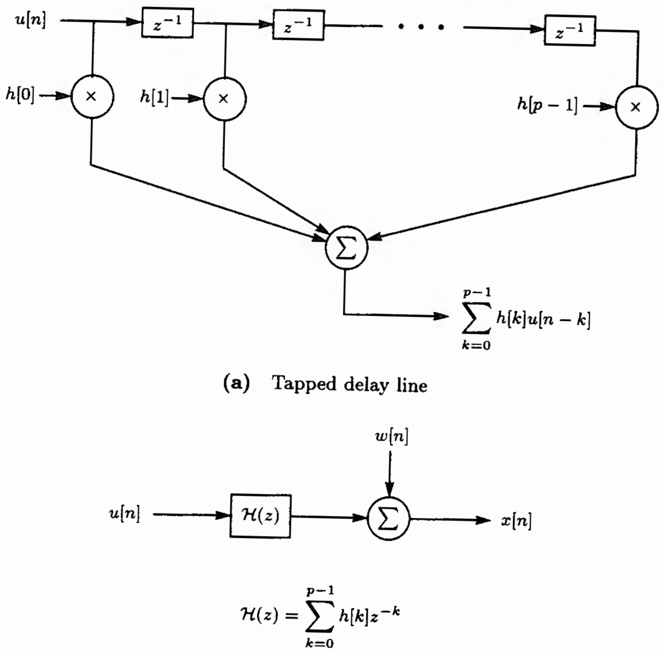
$$\begin{aligned}\mathbf{C}_{\hat{\theta}} &= \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}\\ &= \sigma^2\left(\frac{N}{2}\mathbf{I}\right)^{-1}\\ &= \frac{2\sigma^2}{N}\mathbf{I}.\end{aligned} \qquad (4.16)$$

Because $\hat{\boldsymbol{\theta}}$ is Gaussian and the covariance matrix is diagonal, the amplitude estimates are independent (see Problem 4.6 for an application to sinusoidal detection).

It is seen from this example that a key ingredient in simplifying the computation of the MVU estimator and its covariance matrix is the orthogonality of the columns of $\mathbf{H}$. Note that this property *does not hold* if the frequencies are arbitrarily chosen. ◇

Example 4.3 - System Identification

It is frequently of interest to be able to identify the model of a system from input/output data. A common model is the tapped delay line (TDL) or finite impulse response (FIR) filter shown in Figure 4.3a. The input $u[n]$ is known and is provided to "probe" the system. Ideally, at the output the sequence $\sum_{k=0}^{p-1} h[k]u[n-k]$ is observed from which we would like to estimate the TDL weights $h[k]$, or equivalently, the impulse response of the FIR filter. In practice, however, the output is corrupted by noise, so that the model in Figure 4.3b is more appropriate. Assume that $u[n]$ is provided for $n = 0, 1, \ldots, N-1$ and that the output is observed over the same interval. We then have


u[n]
z^{-1}
z^{-1}
z^{-1}
h[0]
h[1]
h[p − 1]
×
Σ
$\sum_{k=0}^{p-1} h[k]u[n-k]$
w[n]
u[n]
H(z)
Σ
x[n]
$\mathcal{H}(z) = \sum_{k=0}^{p-1} h[k]z^{-k}$

(a) Tapped delay line

(b) Model for noise-corrupted output data

Figure 4.3 System identification model

$$x[n] = \sum_{k=0}^{p-1} h[k]u[n-k] + w[n] \qquad n = 0, 1, \ldots, N-1 \qquad (4.17)$$

where it is assumed that $u[n] = 0$ for $n < 0$. In matrix form we have

$$\mathbf{x} = \underbrace{\begin{bmatrix} u[0] & 0 & \ldots & 0\\ u[1] & u[0] & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots\\ u[N-1] & u[N-2] & \ldots & u[N-p]\end{bmatrix}}_{\mathbf{H}}\underbrace{\begin{bmatrix} h[0]\\ h[1]\\ \vdots\\ h[p-1]\end{bmatrix}}_{\boldsymbol{\theta}} + \mathbf{w}. \qquad (4.18)$$

Assuming $w[n]$ is WGN, (4.18) is in the form of the linear model, and so the MVU estimator of the impulse response is

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}.$$

The covariance matrix is

$$\mathbf{C}_{\hat{\theta}} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}.$$

A key question in system identification is how to choose the probing signal $u[n]$. It is now shown that the signal should be chosen to be pseudorandom noise (PRN) [MacWilliams and Sloane 1976]. Since the variance of $\hat{\theta}_i$ is

$$\mathrm{var}(\hat{\theta}_i) = \mathbf{e}_i^T\mathbf{C}_{\hat{\theta}}\mathbf{e}_i$$

where $\mathbf{e}_i = [0\,0\ldots 0\,1\,0\ldots 0]^T$ with the 1 occupying the ith place, and $\mathbf{C}_{\hat{\theta}}^{-1}$ can be factored as $\mathbf{D}^T\mathbf{D}$ with $\mathbf{D}$ an invertible $p\times p$ matrix, we can use the Cauchy-Schwarz inequality as follows. Noting that

$$1 = (\mathbf{e}_i^T\mathbf{D}^T\mathbf{D}^{T^{-1}}\mathbf{e}_i)^2$$

we can let $\boldsymbol{\xi}_1 = \mathbf{D}\mathbf{e}_i$ and $\boldsymbol{\xi}_2 = \mathbf{D}^{T^{-1}}\mathbf{e}_i$ to yield the inequality

$$(\boldsymbol{\xi}_1^T\boldsymbol{\xi}_2)^2 \le \boldsymbol{\xi}_1^T\boldsymbol{\xi}_1\boldsymbol{\xi}_2^T\boldsymbol{\xi}_2.$$

Because $\boldsymbol{\xi}_1^T\boldsymbol{\xi}_2 = 1$, we have

$$\begin{aligned} 1 &\le (\mathbf{e}_i^T\mathbf{D}^T\mathbf{D}\mathbf{e}_i)(\mathbf{e}_i^T\mathbf{D}^{-1}\mathbf{D}^{T^{-1}}\mathbf{e}_i) \\ &= (\mathbf{e}_i^T\mathbf{C}_{\hat{\theta}}^{-1}\mathbf{e}_i)(\mathbf{e}_i^T\mathbf{C}_{\hat{\theta}}\mathbf{e}_i) \end{aligned}$$

or finally

$$\mathrm{var}(\hat{\theta}_i) \ge \frac{1}{\mathbf{e}_i^T\mathbf{C}_{\hat{\theta}}^{-1}\mathbf{e}_i} = \frac{\sigma^2}{[\mathbf{H}^T\mathbf{H}]_{ii}}.$$

Equality holds or the minimum variance is attained if and only if $\boldsymbol{\xi}_1 = c\boldsymbol{\xi}_2$ for c a constant or

$$\mathbf{D}\mathbf{e}_i = c_i\mathbf{D}^{T^{-1}}\mathbf{e}_i$$

or, equivalently, the conditions for all the variances to be minimized are

$$\mathbf{D}^T\mathbf{D}\mathbf{e}_i = c_i\mathbf{e}_i \qquad i = 1,2,\ldots,p.$$

Noting that

$$\mathbf{D}^T\mathbf{D} = \mathbf{C}_{\hat{\theta}}^{-1} = \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}$$

we have

$$\frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\mathbf{e}_i = c_i\mathbf{e}_i.$$

If we combine these equations in matrix form, then the conditions for achieving the minimum possible variances are

$$\mathbf{H}^T\mathbf{H} = \sigma^2\begin{bmatrix} c_1 & 0 & \ldots & 0 \\ 0 & c_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & c_p \end{bmatrix}.$$

It is now clear that to minimize the variance of the MVU estimator, $u[n]$ should be chosen to make $\mathbf{H}^T\mathbf{H}$ diagonal. Since $[\mathbf{H}]_{ij} = u[i-j]$

$$[\mathbf{H}^T\mathbf{H}]_{ij} = \sum_{n=1}^{N} u[n-i]u[n-j] \qquad i = 1,2,\ldots,p; j = 1,2,\ldots,p \tag{4.19}$$

and for N large we have (see Problem 4.7)

$$[\mathbf{H}^T\mathbf{H}]_{ij} \approx \sum_{n=0}^{N-1-|i-j|} u[n]u[n+|i-j|] \tag{4.20}$$

which can be recognized as a correlation function of the deterministic sequence $u[n]$. Also, with this approximation $\mathbf{H}^T\mathbf{H}$ becomes a symmetric Toeplitz autocorrelation matrix

$$\boxed{\mathbf{H}^T\mathbf{H} = N\begin{bmatrix} r_{uu}[0] & r_{uu}[1] & \ldots & r_{uu}[p-1] \\ r_{uu}[1] & r_{uu}[0] & \ldots & r_{uu}[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{uu}[p-1] & r_{uu}[p-2] & \ldots & r_{uu}[0] \end{bmatrix}}$$

where

$$r_{uu}[k] = \frac{1}{N}\sum_{n=0}^{N-1-k} u[n]u[n+k]$$

may be viewed as an autocorrelation function of $u[n]$. For $\mathbf{H}^T\mathbf{H}$ to be diagonal we require

$$r_{uu}[k] = 0 \qquad k \ne 0$$

which is approximately realized if we use a PRN sequence as our input signal. Finally, under these conditions $\mathbf{H}^T\mathbf{H} = Nr_{uu}[0]\mathbf{I}$, and hence

$$\mathrm{var}(\hat{h}[i]) = \frac{1}{Nr_{uu}[0]/\sigma^2} \qquad i = 0,1,\ldots,p-1 \tag{4.21}$$

and the TDL weight estimators are independent.

As a final consequence of choosing a PRN sequence, we obtain as our MVU estimator

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

where $\mathbf{H}^T\mathbf{H} = Nr_{uu}[0]\mathbf{I}$. Hence, we have

$$\begin{aligned}\hat{h}[i] &= \frac{1}{Nr_{uu}[0]}\sum_{n=0}^{N-1} u[n-i]x[n] \\ &= \frac{\frac{1}{N}\sum_{n=0}^{N-1-i} u[n]x[n+i]}{r_{uu}[0]}\end{aligned} \tag{4.22}$$

since $u[n] = 0$ for $n < 0$. The numerator in (4.22) is just the crosscorrelation $r_{ux}[i]$ between the input and output sequences. Hence, if a PRN input is used to identify the system, then the approximate (for large N) MVU estimator is

$$\hat{h}[i] = \frac{r_{ux}[i]}{r_{uu}[0]} \qquad i = 0, 1, \ldots, p-1 \tag{4.23}$$

where

$$\boxed{\begin{aligned} r_{ux}[i] &= \frac{1}{N}\sum_{n=0}^{N-1-i} u[n]x[n+i] \\ r_{uu}[0] &= \frac{1}{N}\sum_{n=0}^{N-1} u^2[n]. \end{aligned}}$$

◇

See also Problem 4.8 for a spectral interpretation of the system identification problem.

4.5 Extension to the Linear Model

A more general form of the linear model allows for noise that is not white. The *general linear model* assumes that

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

where $\mathbf{C}$ is not necessarily a scaled identity matrix. To determine the MVU estimator, we can repeat the derivation in Section 4.3 (see Problem 4.9). Alternatively, we can use a *whitening* approach as follows. Since $\mathbf{C}$ is assumed to be positive definite, $\mathbf{C}^{-1}$ is positive definite and so can be factored as

$$\mathbf{C}^{-1} = \mathbf{D}^T\mathbf{D} \tag{4.24}$$

where $\mathbf{D}$ is an $N \times N$ invertible matrix. The matrix $\mathbf{D}$ acts as a whitening transformation when applied to $\mathbf{w}$ since

$$\begin{aligned} E\left[(\mathbf{Dw})(\mathbf{Dw})^T)\right] &= \mathbf{DCD}^T \\ &= \mathbf{DD}^{-1}\mathbf{D}^{T^{-1}}\mathbf{D}^T = \mathbf{I}. \end{aligned}$$

=D E[WW^T] D^T = DCD^T

As a consequence, if we transform our generalized model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

to

$$\begin{aligned} \mathbf{x}' &= \mathbf{Dx} \\ &= \mathbf{DH}\boldsymbol{\theta} + \mathbf{Dw} \\ &= \mathbf{H}'\boldsymbol{\theta} + \mathbf{w}' \end{aligned}$$

the noise will be whitened since $\mathbf{w}' = \mathbf{Dw} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the usual linear model will result. The MVU estimator of $\boldsymbol{\theta}$ is, from (4.9),

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\mathbf{H}'^T\mathbf{H}')^{-1}\mathbf{H}'^T\mathbf{x}' \\ &= (\mathbf{H}^T\mathbf{D}^T\mathbf{DH})^{-1}\mathbf{H}^T\mathbf{D}^T\mathbf{Dx} \end{aligned}$$

so that

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}. \tag{4.25}$$

In a similar fashion we find that

$$\mathbf{C}_{\hat{\theta}} = (\mathbf{H}'^T\mathbf{H}')^{-1}$$

or finally

$$\mathbf{C}_{\hat{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}. \tag{4.26}$$

Of course, if $\mathbf{C} = \sigma^2\mathbf{I}$, we have our previous results. The use of the general linear model is illustrated by an example.

Example 4.4 - DC Level in Colored Noise

We now extend Example 3.3 to the colored noise case. If $x[n] = A + w[n]$ for $n = 0, 1, \ldots, N-1$, where $w[n]$ is *colored* Gaussian noise with $N \times N$ covariance matrix $\mathbf{C}$, it immediately follows from (4.25) that with $\mathbf{H} = \mathbf{1} = [1\,1\ldots1]^T$, the MVU estimator of the DC level is

$$\begin{aligned} \hat{A} &= (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} \\ &= \frac{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{x}}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}} \end{aligned}$$

and the variance is, from (4.26),

$$\begin{aligned} \operatorname{var}(\hat{A}) &= (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1} \\ &= \frac{1}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}}. \end{aligned}$$

If $\mathbf{C} = \sigma^2\mathbf{I}$, we have as our MVU estimator the sample mean with a variance of σ^2/N. An interesting interpretation of the MVU estimator follows by considering the factorization of $\mathbf{C}^{-1}$ as $\mathbf{D}^T\mathbf{D}$. We noted previously that $\mathbf{D}$ is a whitening matrix. The MVU estimator is expressed as

$$\begin{aligned}\hat{A} &= \frac{\mathbf{1}^T\mathbf{D}^T\mathbf{D}\mathbf{x}}{\mathbf{1}^T\mathbf{D}^T\mathbf{D}\mathbf{1}} \\ &= \frac{(\mathbf{D}\mathbf{1})^T\mathbf{x}'}{\mathbf{1}^T\mathbf{D}^T\mathbf{D}\mathbf{1}} \\ &= \sum_{n=0}^{N-1} d_n x'[n] \end{aligned} \tag{4.27}$$

where $d_n = [\mathbf{D1}]_n/\mathbf{1}^T\mathbf{D}^T\mathbf{D1}$. According to (4.27), the data are first prewhitened to form $x'[n]$ and then "averaged" using prewhitened averaging weights d_n. The prewhitening has the effect of decorrelating and equalizing the variances of the noises at each observation time before averaging (see Problems 4.10 and 4.11).

◇

Another extension to the linear model allows for signal components that are known. Assume that $\mathbf{s}$ represents a known signal contained in the data. Then, a linear model that incorporates this signal is

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{s} + \mathbf{w}.$$

To determine the MVU estimator let $\mathbf{x}' = \mathbf{x} - \mathbf{s}$, so that

$$\mathbf{x}' = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

which is now in the form of the linear model. The MVU estimator follows as

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T(\mathbf{x} - \mathbf{s}) \tag{4.28}$$

with covariance

$$\mathbf{C}_{\hat{\theta}} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}. \tag{4.29}$$

Example 4.5 - DC Level and Exponential in White Noise

If $x[n] = A + r^n + w[n]$ for $n = 0, 1, \ldots, N-1$, where r is *known*, A is to be estimated, and $w[n]$ is WGN, the model is

$$\mathbf{x} = A\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} + \mathbf{s} + \mathbf{w}$$

where $\mathbf{s} = [1\, r \ldots r^{N-1}]^T$. The MVU estimator, is from (4.28),

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1}(x[n] - r^n)$$

with variance, from (4.29), as

$$\mathrm{var}(\hat{A}) = \frac{\sigma^2}{N}.$$

◇

It should be clear that the two extensions described can be combined to produce the general linear model summarized by the following theorem.

Theorem 4.2 (Minimum Variance Unbiased Estimator for General Linear Model) *If the data can be modeled as*

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{s} + \mathbf{w} \tag{4.30}$$

where $\mathbf{x}$ *is an* $N \times 1$ *vector of observations,* $\mathbf{H}$ *is a known* $N \times p$ *observation matrix* $(N > p)$ *of rank* p, $\boldsymbol{\theta}$ *is a* $p \times 1$ *vector of parameters to be estimated,* $\mathbf{s}$ *is an* $N \times 1$ *vector of known signal samples, and* $\mathbf{w}$ *is an* $N \times 1$ *noise vector with PDF* $\mathcal{N}(\mathbf{0}, \mathbf{C})$, *then the MVU estimator is*

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}(\mathbf{x} - \mathbf{s}) \tag{4.31}$$

and the covariance matrix is

$$\mathbf{C}_{\hat{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}. \tag{4.32}$$

For the general linear model the MVU estimator is efficient in that it attains the CRLB.

This theorem is quite powerful in practice since many signal processing problems can be modeled by (4.30).

References

Graybill, F.A., *Theory and Application of the Linear Model,* Duxbury Press, North Scituate, Mass., 1976.

MacWilliams, F.J., N.J. Sloane, "Pseudo-Random Sequences and Arrays," *Proc. IEEE*, Vol. 64, pp. 1715–1729, Dec. 1976.

Problems

4.1 We wish to estimate the amplitudes of exponentials in noise. The observed data are

$$x[n] = \sum_{i=1}^{p} A_i r_i^n + w[n] \qquad n = 0, 1, \ldots, N-1$$

5.3 Sufficient Statistics

In a previous chapter we found that for the problem of estimating a DC level A in WGN (see Example 3.3), the sample mean

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

was the MVU estimator, having minimum variance σ^2/N. If, on the other hand, we had chosen

$$\check{A} = x[0]$$

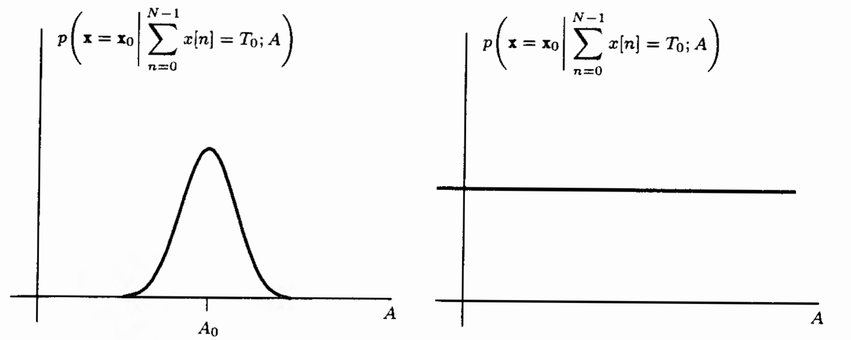
as our estimator, it is immediately clear that even though $\check{A}$ is unbiased, its variance is much larger (being σ^2) than the minimum. Intuitively, the poor performance is a direct result of discarding the data points $\{x[1], x[2], \ldots, x[N-1]\}$ which carry information about A. A reasonable question to ask is Which data samples are pertinent to the estimation problem? or Is there a set of data that is sufficient?. The following data sets may be claimed to be sufficient in that they may be used to compute $\hat{A}$.

$$\begin{aligned} S_1 &= \{x[0], x[1], \ldots, x[N-1]\} \\ S_2 &= \{x[0]+x[1], x[2], x[3], \ldots, x[N-1]\} \\ S_3 &= \left\{\sum_{n=0}^{N-1} x[n]\right\}. \end{aligned}$$

S_1 represents the original data set, which as expected, is always sufficient for the problem. S_2 and S_3 are also sufficient. It is obvious that for this problem there are many sufficient data sets. The data set that contains the least number of elements is called the *minimal* one. If we now think of the elements of these sets as statistics, we say that the N statistics of S_1 are *sufficient* as well as the $(N-1)$ statistics of S_2 and the single statistic of S_3. This latter statistic, $\sum_{n=0}^{N-1} x[n]$, in addition to being a *sufficient statistic*, is the *minimal sufficient statistic*. For estimation of A, once we know $\sum_{n=0}^{N-1} x[n]$, we no longer need the individual data values since all information has been summarized in the sufficient statistic. To quantify what we mean by this, consider the PDF of the data

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \tag{5.1}$$

and assume that $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n] = T_0$ has been observed. Knowledge of the value of this statistic will change the PDF to the conditional one $p(\mathbf{x}|\sum_{n=0}^{N-1} x[n] = T_0; A)$, which now gives the PDF of the observations after the sufficient statistic has been observed. Since the statistic is sufficient for the estimation of A, this conditional PDF should not depend on A. If it did, then we could infer some additional information about A from the data in addition to that already provided by the sufficient statistic. As an example, in Figure 5.1a, if $\mathbf{x} = \mathbf{x}_0$ for an arbitrary $\mathbf{x}_0$, then values of A near A_0 would be more likely. This violates our notion that $\sum_{n=0}^{N-1} x[n]$ is a sufficient statistic. On the other hand, in Figure 5.1b, any value of A is as likely as any other, so that after observing $T(\mathbf{x})$ the data may be discarded. Hence, to verify that a statistic is sufficient we need to determine the conditional PDF and confirm that there is no dependence on A.



(a) Observations provide information after $T(\mathbf{x})$ observed—$T(\mathbf{x})$ is not sufficient

(b) No information from observations after $T(\mathbf{x})$ observed—$T(\mathbf{x})$ is sufficient

Figure 5.1 Sufficient statistic definition

Example 5.1 - Verification of a Sufficient Statistic

Consider the PDF of (5.1). To prove that $\sum_{n=0}^{N-1} x[n]$ is a sufficient statistic we need to determine $p(\mathbf{x}|T(\mathbf{x}) = T_0; A)$, where $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$. By the definition of the conditional PDF we have

$$p(\mathbf{x}|T(\mathbf{x}) = T_0; A) = \frac{p(\mathbf{x}, T(\mathbf{x}) = T_0; A)}{p(T(\mathbf{x}) = T_0; A)}.$$

But note that $T(\mathbf{x})$ is functionally dependent on $\mathbf{x}$, so that the *joint* PDF $p(\mathbf{x}, T(\mathbf{x}) = T_0; A)$ takes on nonzero values only when $\mathbf{x}$ satisfies $T(\mathbf{x}) = T_0$. The joint PDF is therefore $p(\mathbf{x}; A)\delta(T(\mathbf{x}) - T_0)$, where δ is the Dirac delta function (see also Appendix 5A for a further discussion). Thus, we have that

$$p(\mathbf{x}|T(\mathbf{x}) = T_0; A) = \frac{p(\mathbf{x}; A)\delta(T(\mathbf{x}) - T_0)}{p(T(\mathbf{x}) = T_0; A)}. \tag{5.2}$$

Clearly, $T(\mathbf{x}) \sim \mathcal{N}(NA, N\sigma^2)$, so that

$$\begin{aligned} & p(\mathbf{x}; A)\delta(T(\mathbf{x}) - T_0) \\ = & \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]\delta(T(\mathbf{x})-T_0) \\ = & \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{n=0}^{N-1}x^2[n]-2AT(\mathbf{x})+NA^2\right)\right]\delta(T(\mathbf{x})-T_0) \\ = & \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{n=0}^{N-1}x^2[n]-2AT_0+NA^2\right)\right]\delta(T(\mathbf{x})-T_0). \end{aligned}$$

From (5.2) we have

$$\begin{aligned} & p(\mathbf{x}|T(\mathbf{x}) = T_0; A) \\ = & \frac{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]\exp\left[-\frac{1}{2\sigma^2}(-2AT_0+NA^2)\right]}{\frac{1}{\sqrt{2\pi N\sigma^2}}\exp\left[-\frac{1}{2N\sigma^2}(T_0-NA)^2\right]}\delta(T(\mathbf{x})-T_0) \\ = & \frac{\sqrt{N}}{(2\pi\sigma^2)^{\frac{N-1}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]\exp\left[\frac{T_0^2}{2N\sigma^2}\right]\delta(T(\mathbf{x})-T_0) \end{aligned}$$

which as claimed does not depend on A. Therefore, we can conclude that $\sum_{n=0}^{N-1} x[n]$ is a sufficient statistic for the estimation of A. ◇

This example indicates the procedure for verifying that a statistic is sufficient. For many problems the task of evaluating the conditional PDF is formidable, so that an easier approach is needed. Additionally, in Example 5.1 the choice of $\sum_{n=0}^{N-1} x[n]$ for examination as a sufficient statistic was fortuitous. In general an even more difficult problem would be to *identify* potential sufficient statistics. The approach of guessing at a sufficient statistic and then verifying it is, of course, quite unsatisfactory in practice. To alleviate the guesswork we can employ the Neyman-Fisher factorization theorem, which is a simple "turn-the-crank" procedure for finding sufficient statistics.

5.4 Finding Sufficient Statistics

The Neyman-Fisher factorization theorem is now stated, after which we will use it to find sufficient statistics in several examples.

Theorem 5.1 (Neyman-Fisher Factorization) *If we can factor the PDF $p(\mathbf{x};\theta)$ as*

$$p(\mathbf{x};\theta) = g(T(\mathbf{x}),\theta)h(\mathbf{x}) \tag{5.3}$$

where g is a function depending on $\mathbf{x}$ only through $T(\mathbf{x})$ and h is a function depending only on $\mathbf{x}$, then $T(\mathbf{x})$ is a sufficient statistic for θ. Conversely, if $T(\mathbf{x})$ is a sufficient statistic for θ, then the PDF can be factored as in (5.3).

A proof of this theorem is contained in Appendix 5A. It should be mentioned that at times it is not obvious if the PDF can be factored in the required form. If this is the case, then a sufficient statistic may not exist. Some examples are now given to illustrate the use of this powerful theorem.

Example 5.2 - DC Level in WGN

We now reexamine the problem discussed in the previous section. There the PDF was given by (5.1), where we note that σ^2 is assumed known. To demonstrate that a factorization exists we observe that the exponent of the PDF may be rewritten as

$$\sum_{n=0}^{N-1}(x[n]-A)^2 = \sum_{n=0}^{N-1}x^2[n] - 2A\sum_{n=0}^{N-1}x[n] + NA^2$$

so that the PDF is factorable as

$$p(\mathbf{x};A) = \underbrace{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\left(NA^2-2A\sum_{n=0}^{N-1}x[n]\right)\right]}_{g(T(\mathbf{x}),A)}\underbrace{\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]}_{h(\mathbf{x})}.$$

Clearly then, $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$ is a sufficient statistic for A. Note that $T'(\mathbf{x}) = 2\sum_{n=0}^{N-1} x[n]$ is also a sufficient statistic for A, and in fact *any one-to-one function of $\sum_{n=0}^{N-1} x[n]$ is a sufficient statistic* (see Problem 5.12). Hence, sufficient statistics are unique only to within one-to-one transformations. ◇

Example 5.3 - Power of WGN

Now consider the PDF of (5.1) with $A = 0$ and σ^2 as the unknown parameter. Then,

$$p(\mathbf{x};\sigma^2) = \underbrace{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]}_{g(T(\mathbf{x}),\sigma^2)} \cdot \underbrace{1}_{h(\mathbf{x})}.$$

Again it is immediately obvious from the factorization theorem that $T(\mathbf{x}) = \sum_{n=0}^{N-1} x^2[n]$ is a sufficient statistic for σ^2. See also Problem 5.1. ◇

Example 5.4 - Phase of Sinusoid

Recall the problem in Example 3.4 in which we wish to estimate the phase of a sinusoid embedded in WGN or

$$x[n] = A\cos(2\pi f_0 n + \phi) + w[n] \qquad n = 0, 1, \ldots, N-1.$$

Here, the amplitude A and frequency f_0 of the sinusoid are known, as is the noise variance σ^2. The PDF is

$$p(\mathbf{x};\phi) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}[x[n] - A\cos(2\pi f_0 n + \phi)]^2\right\}.$$

The exponent may be expanded as

$$\sum_{n=0}^{N-1} x^2[n] - 2A\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n + \phi) + \sum_{n=0}^{N-1} A^2\cos^2(2\pi f_0 n + \phi)$$

$$= \sum_{n=0}^{N-1} x^2[n] - 2A\left(\sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n\right)\cos\phi$$
$$+2A\left(\sum_{n=0}^{N-1} x[n]\sin 2\pi f_0 n\right)\sin\phi + \sum_{n=0}^{N-1} A^2\cos^2(2\pi f_0 n + \phi).$$

In this example it does not appear that the PDF is factorable as required by the Neyman-Fisher theorem. Hence, no *single* sufficient statistic exists. However, it can be factored as

$$p(\mathbf{x};\phi) =$$
$$\underbrace{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{n=0}^{N-1} A^2\cos^2(2\pi f_0 n + \phi) - 2AT_1(\mathbf{x})\cos\phi + 2AT_2(\mathbf{x})\sin\phi\right]\right\}}_{g(T_1(\mathbf{x}), T_2(\mathbf{x}), \phi)}$$
$$\cdot \underbrace{\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n]\right]}_{h(\mathbf{x})}$$

where

$$\begin{aligned} T_1(\mathbf{x}) &= \sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n \\ T_2(\mathbf{x}) &= \sum_{n=0}^{N-1} x[n]\sin 2\pi f_0 n. \end{aligned}$$

By a slight generalization of the Neyman-Fisher theorem we can conclude that $T_1(\mathbf{x})$ *and* $T_2(\mathbf{x})$ are *jointly* sufficient statistics for the estimation of ϕ. However, no *single* sufficient statistic exists. The reason why we wish to restrict our attention to single sufficient statistics will become clear in the next section. ◇

The concept of jointly sufficient statistics is a simple extension of our previous definition. The r statistics $T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_r(\mathbf{x})$ are jointly sufficient statistics if the conditional PDF $p(\mathbf{x}|T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_r(\mathbf{x}); \theta)$ does not depend on θ. The generalization of the Neyman-Fisher theorem asserts that if $p(\mathbf{x};\theta)$ can be factored as [Kendall and Stuart 1979]

$$p(\mathbf{x};\theta) = g(T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_r(\mathbf{x}), \theta)h(\mathbf{x}) \tag{5.4}$$

then $\{T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_r(\mathbf{x})\}$ are sufficient statistics for θ. It is clear then that the original data are always sufficient statistics since we can let $r = N$ and

$$T_{n+1}(\mathbf{x}) = x[n] \qquad n = 0, 1, \ldots, N-1$$

so that

$$\begin{aligned} g &= p(\mathbf{x};\theta) \\ h &= 1 \end{aligned}$$

and (5.4) holds identically. Of course, they are seldom the *minimal* set of sufficient statistics.

5.5 Using Sufficiency to Find the MVU Estimator

Assuming that we have been able to find a sufficient statistic $T(\mathbf{x})$ for θ, we can make use of the Rao-Blackwell-Lehmann-Scheffe (RBLS) theorem to find the MVU estimator. We will first illustrate the approach with an example and then state the theorem formally.

Example 5.5 - DC Level in WGN

We will continue Example 5.2. Although we already know that $\hat{A} = \bar{x}$ is the MVU estimator (since it is efficient), we will use the RBLS theorem, which can be used even when an efficient estimator does not exist and hence the CRLB method is no longer viable. The procedure for finding the MVU estimator $\hat{A}$ may be implemented in two different ways. They are both based on the sufficient statistic $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$.

1. Find *any* unbiased estimator of A, say $\check{A} = x[0]$, and determine $\hat{A} = E(\check{A}|T)$. The expectation is taken with respect to $p(\check{A}|T)$.

2. Find some function g so that $\hat{A} = g(T)$ is an unbiased estimator of A.

For the first approach we can let the unbiased estimator be $\check{A} = x[0]$ and determine $\hat{A} = E(x[0]|\sum_{n=0}^{N-1} x[n])$. To do so we will need some properties of the conditional

Gaussian PDF. For $[x\, y]^T$ a Gaussian random vector with mean vector $\boldsymbol{\mu} = [E(x)\, E(y)]^T$ and covariance matrix

$$\mathbf{C} = \begin{bmatrix} \mathrm{var}(x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(y,x) & \mathrm{var}(y) \end{bmatrix},$$

it may be shown that (see Appendix 10A)

$$\begin{aligned} E(x|y) &= \int_{-\infty}^{\infty} x p(x|y)\, dx \\ &= \int_{-\infty}^{\infty} x \frac{p(x,y)}{p(y)}\, dx \\ &= E(x) + \frac{\mathrm{cov}(x,y)}{\mathrm{var}(y)}(y - E(y)). \end{aligned} \tag{5.5}$$

Applying this result, we let $x = x[0]$ and $y = \sum_{n=0}^{N-1} x[n]$ and note that

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x[0] \\ \sum_{n=0}^{N-1} x[n] \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}}_{\mathbf{L}} \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}.$$

Hence, the PDF of $[x\, y]^T$ is $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ since this represents a linear transformation of a Gaussian vector, where

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{L}E(\mathbf{x}) = \mathbf{L}A\mathbf{1} = \begin{bmatrix} A \\ NA \end{bmatrix} \\ \mathbf{C} &= \sigma^2 \mathbf{L}\mathbf{L}^T = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & N \end{bmatrix}. \end{aligned}$$

Hence, we have finally from (5.5) that

$$\begin{aligned} \hat{A} &= E(x|y) = A + \frac{\sigma^2}{N\sigma^2}\left(\sum_{n=0}^{N-1} x[n] - NA\right) \\ &= \frac{1}{N}\sum_{n=0}^{N-1} x[n] \end{aligned}$$

which is the MVU estimator. This approach, requiring evaluation of a conditional expectation, is usually mathematically intractable.

Turning our attention to the second approach, we need to find some function g so that

$$\hat{A} = g\left(\sum_{n=0}^{N-1} x[n]\right)$$

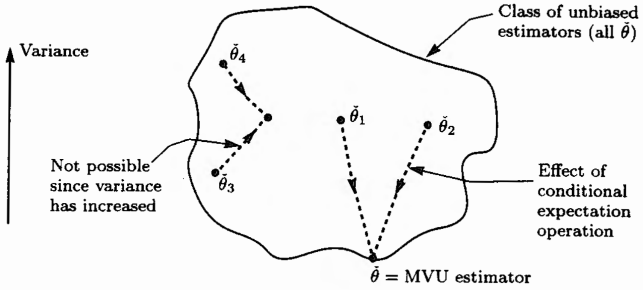

Variance
Class of unbiased estimators (all $\check{\theta}$)
$\check{\theta}_4$
$\check{\theta}_1$
$\check{\theta}_2$
$\check{\theta}_3$
Not possible since variance has increased
Effect of conditional expectation operation
$\hat{\theta}$ = MVU estimator

Figure 5.2 RBLS argument for MVU estimator

is an unbiased estimator of A. By inspection this is $g(x) = x/N$, which yields

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

as the MVU estimator. This alternative method is much easier to apply, and therefore in practice, it is the one we generally employ. ◇

We now formally state the RBLS theorem.

Theorem 5.2 (Rao-Blackwell-Lehmann-Scheffe) *If $\check{\theta}$ is an unbiased estimator of θ and $T(\mathbf{x})$ is a sufficient statistic for θ, then $\hat{\theta} = E(\check{\theta}|T(\mathbf{x}))$ is*

1. a valid estimator for θ (not dependent on θ)

2. unbiased

3. of lesser or equal variance than that of $\check{\theta}$, for all θ.

Additionally, if the sufficient statistic is complete, then $\hat{\theta}$ is the MVU estimator.

A proof is given in Appendix 5B. In the previous example we saw that $E(x[0]|\sum_{n=0}^{N-1} x[n]) = \bar{x}$ did not depend on A, making it a valid estimator, was unbiased, and had less variance than $x[0]$. That there is no other estimator with less variance, as Theorem 5.2 asserts, follows from the property that the sufficient statistic $\sum_{n=0}^{N-1} x[n]$ is a *complete* sufficient statistic. In essence, a statistic is complete if *there is only one function of the statistic that is unbiased.* The argument that $\hat{\theta} = E(\check{\theta}|T(\mathbf{x}))$ is the MVU estimator is now given. Consider all possible unbiased estimators of θ, as depicted in Figure 5.2. By determining $E(\check{\theta}|T(\mathbf{x}))$ we can lower the variance of the estimator (property 3 of Theorem 5.2) and still remain within the class (property 2 of Theorem 5.2). But $E(\check{\theta}|T(\mathbf{x}))$ is solely a function of the sufficient statistic $T(\mathbf{x})$ since

$$\begin{aligned} \hat{\theta} &= E(\check{\theta}|T(\mathbf{x})) = \int \check{\theta} p(\check{\theta}|T(\mathbf{x}))\, d\check{\theta} \\ &= g(T(\mathbf{x})). \end{aligned} \tag{5.6}$$

If $T(\mathbf{x})$ is complete, there is but *one* function of T that is an unbiased estimator. Hence, $\hat{\theta}$ is *unique*, independent of the $\check{\theta}$ we choose from the class shown in Figure 5.2. Every $\check{\theta}$ maps into the *same estimator* $\hat{\theta}$. Because the variance of $\hat{\theta}$ must be less than $\check{\theta}$ for any $\check{\theta}$ within the class (property 3 of Theorem 5.2), we conclude that $\hat{\theta}$ must be the MVU estimator. In summary, the MVU estimator can be found by taking *any* unbiased estimator and carrying out the operations in (5.6). Alternatively, since there is only one function of the sufficient statistic that leads to an unbiased estimator, we need only find the unique g to make the sufficient statistic unbiased. For this latter approach we found in Example 5.5 that $g(\sum_{n=0}^{N-1} x[n]) = \sum_{n=0}^{N-1} x[n]/N$.

The property of completeness depends on the PDF of $\mathbf{x}$, which in turn determines the PDF of the sufficient statistic. For many practical cases of interest it holds. In particular, for the exponential family of PDFs (see Problems 5.14 and 5.15) this condition is satisfied. To validate that a sufficient statistic is complete is in general quite difficult, and we refer the reader to the discussions in [Kendall and Stuart 1979]. A flavor for the concept of completeness is provided by the next two examples.

Example 5.6 - Completeness of a Sufficient Statistic

For the estimation of A, the sufficient statistic $\sum_{n=0}^{N-1} x[n]$ is complete or there is but one function g for which $E[g(\sum_{n=0}^{N-1} x[n])] = A$. Suppose, however, that there exists a second function h for which $E[h(\sum_{n=0}^{N-1} x[n])] = A$. Then, it would follow that with $T = \sum_{n=0}^{N-1} x[n]$,

$$E\left[g(T) - h(T)\right] = A - A = 0 \qquad \text{for all } A$$

or since $T \sim \mathcal{N}(NA, N\sigma^2)$

$$\int_{-\infty}^{\infty} v(T) \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left[-\frac{1}{2N\sigma^2}(T - NA)^2\right] dT = 0 \qquad \text{for all } A$$

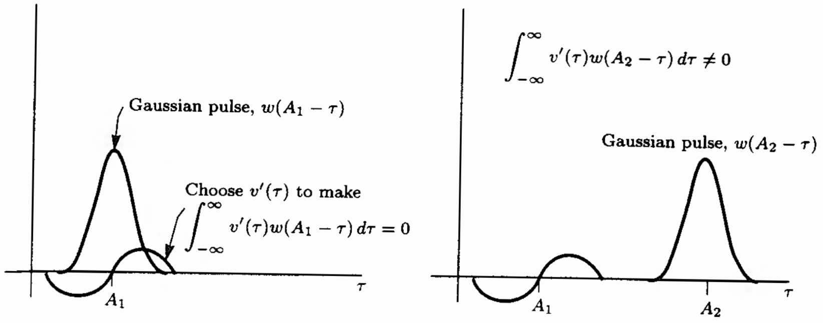
where $v(T) = g(T) - h(T)$. Letting $\tau = T/N$ and $v'(\tau) = v(N\tau)$, we have

$$\int_{-\infty}^{\infty} v'(\tau) \frac{N}{\sqrt{2\pi N\sigma^2}} \exp\left[-\frac{N}{2\sigma^2}(A - \tau)^2\right] d\tau = 0 \qquad \text{for all } A \qquad (5.7)$$

which may be recognized as the convolution of a function $v'(\tau)$ with a Gaussian pulse $w(\tau)$ (see Figure 5.3). For the result to be zero for all A, $v'(\tau)$ must be identically zero. To see this recall that a signal is zero if and only if its Fourier transform is identically zero, resulting in the condition

$$V'(f)W(f) = 0 \qquad \text{for all } f$$

where $V'(f) = \mathcal{F}\{v'(\tau)\}$ and $W(f)$ is the Fourier transform of the Gaussian pulse in (5.7). Since $W(f)$ is also Gaussian and therefore positive for all f, we have that the condition is satisfied if and only if $V'(f) = 0$ for all f. Hence, we must have that $v'(\tau) = 0$ for all τ. This implies that $g = h$ or that the function g is unique. ◇


Gaussian pulse, $w(A_1 - \tau)$
Choose $v'(\tau)$ to make
$\int_{-\infty}^{\infty} v'(\tau)w(A_1 - \tau)\,d\tau = 0$
A_1
τ
$\int_{-\infty}^{\infty} v'(\tau)w(A_2 - \tau)\,d\tau \neq 0$
Gaussian pulse, $w(A_2 - \tau)$
A_1
A_2
τ

(a) Integral equals zero **(b)** Integral does not equal zero

Figure 5.3 Completeness condition for sufficient statistic (satisfied)

Example 5.7 - Incomplete Sufficient Statistic

Consider the estimation of A for the datum

$$x[0] = A + w[0]$$

where $w[0] \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$. A sufficient statistic is $x[0]$, being the only available data, and furthermore, $x[0]$ is an unbiased estimator of A. We may conclude that $g(x[0]) = x[0]$ is a viable candidate for the MVU estimator. That it is actually the MVU estimator still requires us to verify that it is a complete sufficient statistic. As in the previous example, we suppose that there exists another function h with the unbiased property $h(x[0]) = A$ and attempt to prove that $h = g$. Again letting $v(T) = g(T) - h(T)$, we examine the possible solutions for v of the equation

$$\int_{-\infty}^{\infty} v(T)p(\mathbf{x}; A)\,d\mathbf{x} = 0 \qquad \text{for all } A.$$

For this problem, however, $\mathbf{x} = x[0] = T$, so that

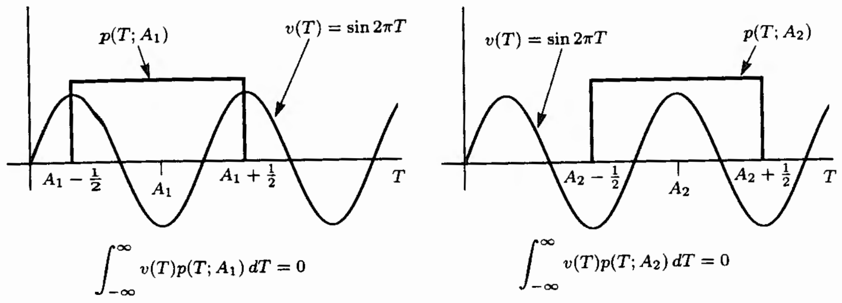
$$\int_{-\infty}^{\infty} v(T)p(T; A)\,dT = 0 \qquad \text{for all A.}$$

But

$$p(T; A) = \begin{cases} 1 & A - \frac{1}{2} \leq T \leq A + \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

so that the condition reduces to

$$\int_{A-\frac{1}{2}}^{A+\frac{1}{2}} v(T)\,dT = 0 \qquad \text{for all } A.$$



(a) Integral equals zero (b) Integral equals zero

Figure 5.4 Completeness condition for sufficient statistic (not satisfied)

The nonzero function $v(T)=\sin 2\pi T$ will satisfy this condition as illustrated in Figure 5.4. Hence a solution is

$$v(T)=g(T)-h(T)=\sin 2\pi T$$

or

$$h(T)=T-\sin 2\pi T.$$

As a result, the estimator

$$\hat{A}=x[0]-\sin 2\pi x[0]$$

is also based on the sufficient statistic and is unbiased for A. Having found at least one other unbiased estimator that is also a function of the sufficient statistic, we may conclude that the sufficient statistic is not complete. The RBLS theorem no longer holds, and it is not possible to assert that $\hat{A}=x[0]$ is the MVU estimator. $\diamond$

To summarize the completeness condition, we say that a sufficient statistic is complete if the condition

$$\int_{-\infty}^{\infty} v(T)p(T;\theta)\,dT=0 \qquad \text{for all } \theta \tag{5.8}$$

is *satisfied only by the zero function* or by $v(T)=0$ for all T.

At this point it is worthwhile to review our results and then apply them to an estimation problem for which we do not know the MVU estimator. The procedure is as follows (see also Figure 5.5):

1. Find a single sufficient statistic for θ, that is, $T(\mathbf{x})$, by using the Neyman-Fisher factorization theorem.

2. Determine if the sufficient statistic is complete and, if so, proceed; if not, this approach cannot be used.

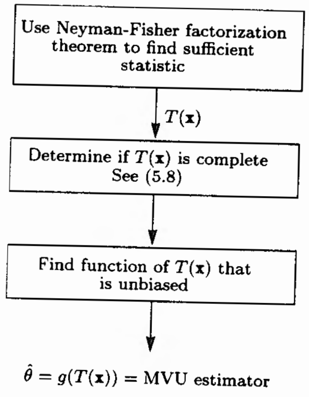

Use Neyman-Fisher factorization theorem to find sufficient statistic
$T(\mathbf{x})$
Determine if $T(\mathbf{x})$ is complete See (5.8)
Find function of $T(\mathbf{x})$ that is unbiased
$\hat{\theta}=g(T(\mathbf{x}))=$ MVU estimator

Figure 5.5 Procedure for finding MVU estimator (scalar parameter)

3. Find a function g of the sufficient statistic that yields an unbiased estimator $\hat{\theta}=g(T(\mathbf{x}))$. The MVU estimator is then $\hat{\theta}$.

As an alternative implementation of step 3 we may

3.′ Evaluate $\hat{\theta}=E(\check{\theta}|T(\mathbf{x}))$, where $\check{\theta}$ is any unbiased estimator.

However, in practice the conditional expectation evaluation is usually too tedious. The next example illustrates the overall procedure.

Example 5.8 - Mean of Uniform Noise

We observe the data

$$x[n]=w[n] \qquad n=0,1,\ldots,N-1$$

where $w[n]$ is IID noise with PDF $\mathcal{U}[0,\beta]$ for $\beta>0$. We wish to find the MVU estimator for the mean $\theta=\beta/2$. Our initial approach of using the CRLB for finding an efficient and hence MVU estimator cannot even be tried for this problem. This is because the PDF does not satisfy the required regularity conditions (see Problem 3.1). A natural estimator of θ is the sample mean or

$$\hat{\theta}=\frac{1}{N}\sum_{n=0}^{N-1}x[n].$$

The sample mean is easily shown to be unbiased and to have variance

$$\begin{aligned}\operatorname{var}(\hat{\theta}) &= \frac{1}{N}\operatorname{var}(x[n])\\ &= \frac{\beta^2}{12N}.\end{aligned} \tag{5.9}$$

6.3 Definition of the BLUE

We observe the data set $\{x[0], x[1], \ldots, x[N-1]\}$ whose PDF $p(\mathbf{x}; \theta)$ depends on an unknown parameter θ. The BLUE restricts the estimator to be linear in the data or

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] \tag{6.1}$$

where the a_n's are constants yet to be determined. (See also Problem 6.6 for a similar definition except for the addition of a constant.) Depending on the a_n's chosen, we may generate a large number of different estimators of θ. However, the best estimator or BLUE is defined to be the one that is unbiased and has minimum variance. Before determining the a_n's that yield the BLUE, some comments about the optimality of the BLUE are in order. Since we are restricting the class of estimators to be linear, the BLUE will be optimal (that is to say, the MVU estimator) only when the MVU estimator turns out to be linear. For example, for the problem of estimating the value of a DC level in WGN (see Example 3.3) the MVU estimator is the sample mean

$$\hat{\theta} = \bar{x} = \sum_{n=0}^{N-1} \frac{1}{N} x[n]$$

which is clearly linear in the data. Hence, if we restrict our attention to only linear estimators, then we will lose nothing since the MVU estimator is within this class. Figure 6.1a depicts this idea. On the other hand, for the problem of estimating the mean of uniformly distributed noise (see Example 5.8), the MVU estimator was found to be

$$\hat{\theta} = \frac{N+1}{2N} \max x[n]$$

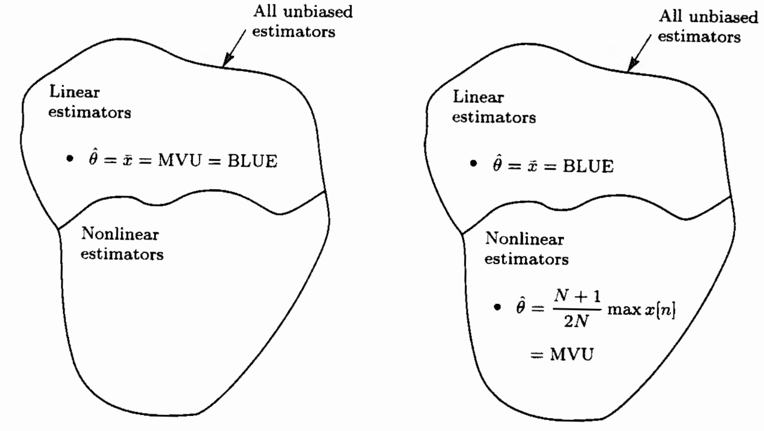
which is nonlinear in the data. If we restrict our estimator to be linear, then the BLUE is the sample mean, as we will see shortly. The BLUE for this problem is suboptimal, as illustrated in Figure 6.1b. As further shown in Example 5.8, the difference in performance is substantial. Unfortunately, without knowledge of the PDF there is no way to determine the loss in performance by resorting to a BLUE.

Finally, for some estimation problems the use of a BLUE can be totally inappropriate. Consider the estimation of the power of WGN. It is easily shown that the MVU estimator is (see Example 3.6)

$$\hat{\sigma^2} = \frac{1}{N} \sum_{n=0}^{N-1} x^2[n]$$

which is nonlinear in the data. If we force the estimator to be linear as per (6.1), so that

$$\hat{\sigma^2} = \sum_{n=0}^{N-1} a_n x[n],$$


All unbiased estimators
Linear estimators
• $\hat{\theta} = \bar{x} = \text{MVU} = \text{BLUE}$
Nonlinear estimators
All unbiased estimators
Linear estimators
• $\hat{\theta} = \bar{x} = \text{BLUE}$
Nonlinear estimators
• $\hat{\theta} = \frac{N+1}{2N} \max x[n]$
$= \text{MVU}$

(a) DC level in WGN; BLUE is optimal

(b) Mean of uniform noise; BLUE is suboptimal

Figure 6.1 Optimality of BLUE

the expected value of the estimator becomes

$$E(\hat{\sigma^2}) = \sum_{n=0}^{N-1} a_n E(x[n]) = 0$$

since $E(x[n]) = 0$ for all n. Thus, we cannot even find a single linear estimator that is unbiased, let alone one that has minimum variance. Although the BLUE is unsuitable for this problem, a BLUE utilizing the *transformed data* $y[n] = x^2[n]$ would produce a viable estimator since for

$$\hat{\sigma^2} = \sum_{n=0}^{N-1} a_n y[n] = \sum_{n=0}^{N-1} a_n x^2[n]$$

the unbiased constraint yields

$$E(\hat{\sigma^2}) = \sum_{n=0}^{N-1} a_n \sigma^2 = \sigma^2.$$

There are many values of the a_n's that could satisfy this constraint. Can you guess what the a_n's should be to yield the BLUE? (See also Problem 6.5 for an example of this data transformation approach.) Hence, with enough ingenuity the BLUE may still be used if the data are first transformed suitably.

6.4 Finding the BLUE

To determine the BLUE we constrain $\hat{\theta}$ to be linear and unbiased and then find the a_n's to minimize the variance. The unbiased constraint, is from (6.1),

$$E(\hat{\theta}) = \sum_{n=0}^{N-1} a_n E(x[n]) = \theta. \tag{6.2}$$

The variance of $\hat{\theta}$ is

$$\mathrm{var}(\hat{\theta}) = E\left[\left(\sum_{n=0}^{N-1} a_n x[n] - E\left(\sum_{n=0}^{N-1} a_n x[n]\right)\right)^2\right].$$

But by using (6.2) and letting $\mathbf{a} = [a_0\, a_1 \ldots a_{N-1}]^T$ we have

$$\begin{aligned} \mathrm{var}(\hat{\theta}) &= E\left[(\mathbf{a}^T\mathbf{x} - \mathbf{a}^T E(\mathbf{x}))^2\right] \\ &= E\left[(\mathbf{a}^T(\mathbf{x} - E(\mathbf{x})))^2\right] \\ &= E\left[\mathbf{a}^T(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T\mathbf{a}\right] \\ &= \mathbf{a}^T\mathbf{C}\mathbf{a}. \end{aligned} \tag{6.3}$$

The vector $\mathbf{a}$ of weights is found by minimizing (6.3) subject to the constraint of (6.2). Before proceeding we need to assume some form for $E(x[n])$. In order to satisfy the unbiased constraint, $E(x[n])$ must be linear in θ or

$$E(x[n]) = s[n]\theta \tag{6.4}$$

where the $s[n]$'s are *known*. Otherwise, it may be impossible to satisfy the constraint. As an example, if $E(x[n]) = \cos\theta$, then the unbiased constraint is $\sum_{n=0}^{N-1} a_n \cos\theta = \theta$ for all θ. Clearly, there do not exist a_n's that will satisfy the unbiased constraint. Note that if we write $x[n]$ as

$$x[n] = E(x[n]) + [x[n] - E(x[n])]$$

then by viewing $x[n] - E(x[n])$ as noise or $w[n]$ we have

$$x[n] = \theta s[n] + w[n].$$

The assumption of (6.4) means that *the BLUE is applicable to amplitude estimation of known signals in noise.* To generalize its use we require a nonlinear transformation of the data as already described.

With the assumption given by (6.4) we now summarize the estimation problem. To find the BLUE we need to minimize the variance

$$\mathrm{var}(\hat{\theta}) = \mathbf{a}^T\mathbf{C}\mathbf{a}$$

subject to the unbiased constraint, which from (6.2) and (6.4) becomes

$$\begin{aligned} \sum_{n=0}^{N-1} a_n E(x[n]) &= \theta \\ \sum_{n=0}^{N-1} a_n s[n]\theta &= \theta \\ \sum_{n=0}^{N-1} a_n s[n] &= 1 \end{aligned}$$

or

$$\mathbf{a}^T\mathbf{s} = 1$$

where $\mathbf{s} = [s[0]\, s[1] \ldots s[N-1]]^T$. The solution to this minimization problem is derived in Appendix 6A as

$$\mathbf{a}_{\mathrm{opt}} = \frac{\mathbf{C}^{-1}\mathbf{s}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}$$

so that the BLUE is

$$\hat{\theta} = \frac{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{x}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}} \tag{6.5}$$

and has minimum variance

$$\mathrm{var}(\hat{\theta}) = \frac{1}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}. \tag{6.6}$$

Note from (6.4) that since $E(\mathbf{x}) = \theta\mathbf{s}$, the BLUE is unbiased

$$\begin{aligned} E(\hat{\theta}) &= \frac{\mathbf{s}^T\mathbf{C}^{-1}E(\mathbf{x})}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}} \\ &= \frac{\mathbf{s}^T\mathbf{C}^{-1}\theta\mathbf{s}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}} \\ &= \theta. \end{aligned}$$

Also, as we asserted earlier, to determine the BLUE we only require knowledge only of

1. $\mathbf{s}$ or the scaled mean

2. $\mathbf{C}$, the covariance

or the first two moments but not the entire PDF. Some examples now follow.

Example 6.1 - DC Level in White Noise

If we observe

$$x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $w[n]$ is white noise with variance σ^2 (*and of unspecified PDF*), then the problem is to estimate A. Since $w[n]$ is not necessarily Gaussian, the noise samples may be

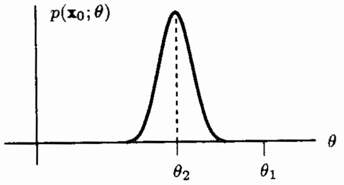


Figure 7.1 Rationale for maximum likelihood estimator

which from (7.2) is the CRLB!

Summarizing our results, the proposed estimator given by (7.6) is asymptotically unbiased and asymptotically achieves the CRLB. Hence, it is *asymptotically efficient*. Furthermore, by the central limit theorem the random variable $\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]$ is Gaussian as $N \to \infty$. Because $\hat{A}$ is a linear function of this Gaussian random variable for large data records (as per (7.7)), it too will have a Gaussian PDF. ◇

The proposed estimator is termed the MLE. How it was found is discussed in the next section.

7.4 Finding the MLE

The MLE for a scalar parameter is defined to be *the value of θ that maximizes $p(\mathbf{x};\theta)$ for $\mathbf{x}$ fixed*, i.e., the value that maximizes the likelihood function. The maximization is performed over the allowable range of θ. In the previous example this was $A > 0$. Since $p(\mathbf{x};\theta)$ will also be a function of $\mathbf{x}$, the maximization produces a $\hat{\theta}$ that is a function of $\mathbf{x}$. The rationale for the MLE hinges on the observation that $p(\mathbf{x};\theta)\,d\mathbf{x}$ gives the probability of observing $\mathbf{x}$ in a small volume for a given θ. In Figure 7.1 the PDF is evaluated for $\mathbf{x} = \mathbf{x}_0$ and then plotted versus θ. The value of $p(\mathbf{x} = \mathbf{x}_0;\theta)\,d\mathbf{x}$ for each θ tells us the probability of observing $\mathbf{x}$ in the region in R^N centered around $\mathbf{x}_0$ with volume $d\mathbf{x}$, assuming the given value of θ. If $\mathbf{x} = \mathbf{x}_0$ had indeed been observed, then inferring that $\theta = \theta_1$ would be unreasonable. Because if $\theta = \theta_1$, the probability of actually observing $\mathbf{x} = \mathbf{x}_0$ would be small. It is more "likely" that $\theta = \theta_2$ is the true value. It yields a high probability of observing $\mathbf{x} = \mathbf{x}_0$, the data that were *actually observed*. Thus, we choose $\hat{\theta} = \theta_2$ as our estimate or the value that maximizes $p(\mathbf{x} = \mathbf{x}_0;\theta)$ over the allowable range of θ. We now continue with our example.

Example 7.3 - DC Level in White Gaussian Noise - Modified (continued)

To actually find the MLE for this problem we first write the PDF from (7.1) as

$$p(\mathbf{x};A) = \frac{1}{(2\pi A)^{\frac{N}{2}}}\exp\left[-\frac{1}{2A}\sum_{n=0}^{N-1}(x[n]-A)^2\right].$$

Considering this as a function of A, it becomes the likelihood function. Differentiating the log-likelihood function, we have

$$\frac{\partial \ln p(\mathbf{x};A)}{\partial A} = -\frac{N}{2A} + \frac{1}{A}\sum_{n=0}^{N-1}(x[n]-A) + \frac{1}{2A^2}\sum_{n=0}^{N-1}(x[n]-A)^2$$

and setting it equal to zero produces

$$\hat{A}^2 + \hat{A} - \frac{1}{N}\sum_{n=0}^{N-1}x^2[n] = 0.$$

Solving for $\hat{A}$ produces the two solutions

$$\hat{A} = -\frac{1}{2} \pm \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}x^2[n] + \frac{1}{4}}.$$

We choose the solution

$$\hat{A} = -\frac{1}{2} + \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}x^2[n] + \frac{1}{4}}$$

to correspond to the permissible range of A or $A > 0$. Note that $\hat{A} > 0$ for all possible values of $\frac{1}{N}\sum_{n=0}^{N-1}x^2[n]$. Finally, that $\hat{A}$ indeed maximizes the log-likelihood function is verified by examining the second derivative. ◇

Not only does the maximum likelihood procedure yield an estimator that is asymptotically efficient, it also sometimes yields an efficient estimator for *finite* data records. This is illustrated in the following example.

Example 7.4 - DC Level in White Gaussian Noise

For the received data

$$x[n] = A + w[n] \qquad n = 0,1,\ldots,N-1$$

where A is the unknown level to be estimated and $w[n]$ is WGN with known variance σ^2, the PDF is

$$p(\mathbf{x};A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right].$$

Taking the derivative of the log-likelihood function produces

$$\frac{\partial \ln p(\mathbf{x};A)}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)$$

which being set equal to zero yields the MLE

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n].$$

But we have already seen that the sample mean is an efficient estimator (see Example 3.3). Hence, the MLE is efficient. ◇

This result is true in general. *If an efficient estimator exists, the maximum likelihood procedure will produce it.* See Problem 7.12 for an outline of the proof.

7.5 Properties of the MLE

The example discussed in Section 7.3 led to an estimator that for large data records (or asymptotically) was unbiased, achieved the CRLB, and had a Gaussian PDF. In summary, the MLE was distributed as

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, I^{-1}(\theta)) \tag{7.8}$$

where $\overset{a}{\sim}$ denotes "asymptotically distributed according to." This result is quite general and forms the basis for claiming optimality of the MLE. Of course, in practice it is seldom known in advance how large N must be in order for (7.8) to hold. An analytical expression for the PDF of the MLE is usually impossible to derive. As an alternative means of assessing performance, a computer simulation is usually required, as discussed next.

Example 7.5 - DC Level in White Gaussian Noise - Modified (continued)

A computer simulation was performed to determine how large the data record had to be for the asymptotic results to apply. In principle the exact PDF of $\hat{A}$ (see (7.6)) could be found but would be extremely tedious. Using the Monte Carlo method (see Appendix 7A), $M = 1000$ realizations of $\hat{A}$ were generated for various data record lengths. The mean $E(\hat{A})$ and variance $\mathrm{var}(\hat{A})$ were *estimated* by

$$\widehat{E(\hat{A})} = \frac{1}{M}\sum_{i=1}^{M}\hat{A}_i \tag{7.9}$$

$$\widehat{\mathrm{var}(\hat{A})} = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{A}_i - \widehat{E(\hat{A})}\right)^2. \tag{7.10}$$

For a value of A equal to 1 the results are shown in Table 7.1 for various data record lengths. Instead of the asymptotic variance or the CRLB of (7.2), we tabulate

$$N\mathrm{var}(\hat{A}) = \frac{A^2}{A+\frac{1}{2}}$$

since this is independent of N, allowing us to check the convergence more readily. The theoretical asymptotic values of the mean and normalized variance are

TABLE 7.1 Theoretical Asymptotic and Actual Mean and Variance for Estimator in Example 7.2

Data Record Length, N	Mean, $E(\hat{A})$	$N\times$Variance, $N\,\mathrm{var}(\hat{A})$
5	0.954	0.624
10	0.976	0.648
15	0.991	0.696
20	0.996 (0.987)	0.707 (0.669)
25	0.994	0.656
Theoretical asymptotic value	1	0.667

$$\begin{aligned} E(\hat{A}) &= A = 1 \\ N\mathrm{var}(\hat{A}) &= \frac{2}{3}. \end{aligned}$$

It is observed from Table 7.1 that the mean converges at about $N = 20$ samples, while the variance jumps around somewhat for $N \geq 15$ samples. The latter is due to the statistical fluctutations in estimating the variance via a computer simulation, as well as possible inaccuracies in the random number generator. To check this the number of realizations was increased to $M = 5000$ for a data record length of $N = 20$. This resulted in the mean and normalized variance shown in parentheses. The normalized variance is now nearly identical to its asymptotic value, whereas for some unknown reason (presumably the random number generator) the mean is off slightly from its asymptotic value. (See also Problem 9.8 for a more accurate formula for $E(\hat{A})$.)

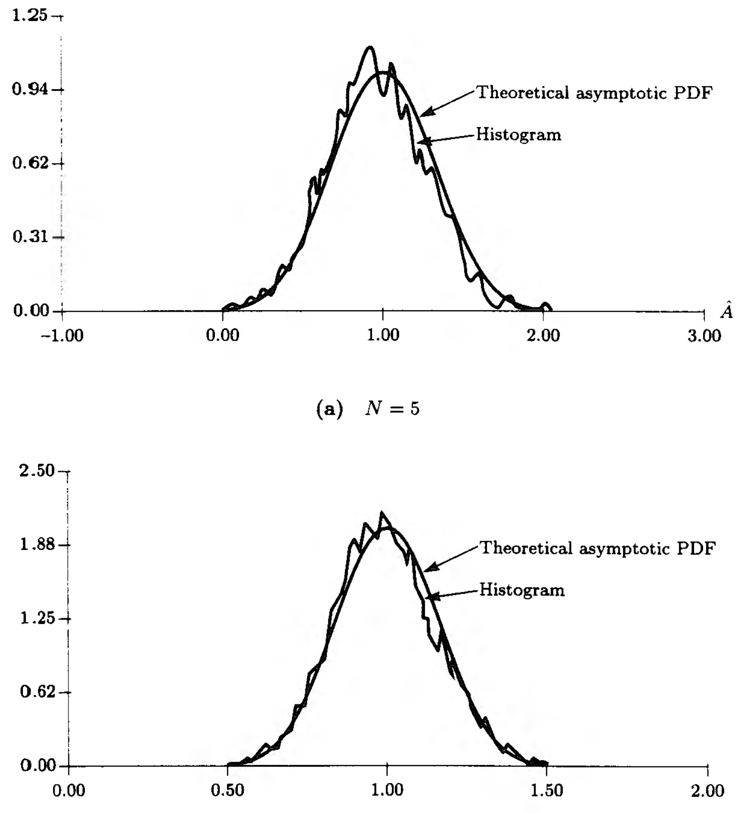
Next, the PDF of $\hat{A}$ was determined using a Monte Carlo computer simulation. This was done for data record lengths of $N = 5$ and $N = 20$. According to (7.8), the asymptotic PDF is

$$\hat{A} \overset{a}{\sim} \mathcal{N}(A, I^{-1}(A)),$$

which for $A = 1$ becomes, upon using (7.2),

$$\hat{A} \overset{a}{\sim} \mathcal{N}(1, \frac{2/3}{N}).$$

In Figure 7.2 the theoretical PDF and estimated PDF or *histogram* (see Appendix 7A) are shown. To construct the histogram we used $M = 5000$ realizations of $\hat{A}$ and divided the horizontal axis into 100 cells or divisions. Note that for $N = 5$ the estimated PDF is somewhat displaced to the left, in accordance with the mean being too small (see Table 7.1). For $N = 20$, however, the match is better, although the estimated PDF still appears to be skewed to the left. Presumably for larger data records the asymptotic PDF will more closely match the true one. ◇



(a) $N = 5$

(b) $N = 20$

Figure 7.2 Theoretical PDF and histogram

In describing the performance of estimators whose PDFs cannot be determined analytically, we must frequently resort to computer simulations, as in the previous example. In practice, the computer has become the dominant tool for analyzing the performance of nonlinear estimators. For this reason it is important to be able to carry out such a simulation. In Appendix 7A a description of the computer methods used in the previous example is given. Of course, the subject of Monte Carlo computer methods for statistical evaluation warrants a more complete discussion. The interested reader should consult the following references: [Bendat and Piersol 1971, Schwartz and Shaw

1975]. Some practice in performing these simulations is provided by Problems 7.13 and 7.14. We now summarize the asymptotic properties of the MLE in a theorem.

Theorem 7.1 (Asymptotic Properties of the MLE) *If the PDF $p(\mathbf{x};\theta)$ of the data $\mathbf{x}$ satisfies some "regularity" conditions, then the MLE of the unknown parameter θ is asymptotically distributed (for large data records) according to*

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, I^{-1}(\theta)) \tag{7.11}$$

where $I(\theta)$ is the Fisher information evaluated at the true value of the unknown parameter.

The regularity conditions require the existence of the derivatives of the log-likelihood function, as well as the Fisher information being nonzero, as described more fully in Appendix 7B. An outline of the proof for IID observations is also given there.

From the asymptotic distribution, the MLE is seen to be asymptotically unbiased and asymptotically attains the CRLB. It is therefore ***asymptotically efficient***, and hence ***asymptotically optimal***. Of course, in practice the key question is always How large does N have to be for the asymptotic properties to apply? Fortunately, for many cases of interest the data record lengths are not excessive, as illustrated by Example 7.5. Another example follows.

Example 7.6 - MLE of the Sinusoidal Phase

We now reconsider the problem in Example 3.4 in which we wish to estimate the phase ϕ of a sinusoid embedded in noise or

$$x[n] = A\cos(2\pi f_0 n + \phi) + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $w[n]$ is WGN with variance σ^2 and the amplitude A and frequency f_0 are assumed to be known. We saw in Chapter 5 that no *single* sufficient statistic exists for this problem. The sufficient statistics were

$$\begin{aligned} T_1(\mathbf{x}) &= \sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n) \\ T_2(\mathbf{x}) &= \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n). \end{aligned} \tag{7.12}$$

The MLE is found by maximizing $p(\mathbf{x};\phi)$ or

$$p(\mathbf{x};\phi) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left(x[n] - A\cos(2\pi f_0 n + \phi)\right)^2\right]$$

or, equivalently, by minimizing

$$J(\phi) = \sum_{n=0}^{N-1}\left(x[n] - A\cos(2\pi f_0 n + \phi)\right)^2. \tag{7.13}$$

Differentiating with respect to ϕ produces

$$\frac{\partial J(\phi)}{\partial \phi} = 2\sum_{n=0}^{N-1}\left(x[n] - A\cos(2\pi f_0 n + \phi)\right) A \sin(2\pi f_0 n + \phi)$$

and setting it equal to zero yields

$$\sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n + \hat{\phi}) = A\sum_{n=0}^{N-1}\sin(2\pi f_0 n + \hat{\phi})\cos(2\pi f_0 n + \hat{\phi}). \tag{7.14}$$

But the right-hand side may be approximated since (see Problem 3.7)

$$\frac{1}{N}\sum_{n=0}^{N-1}\sin(2\pi f_0 n + \hat{\phi})\cos(2\pi f_0 n + \hat{\phi}) = \frac{1}{2N}\sum_{n=0}^{N-1}\sin(4\pi f_0 n + 2\hat{\phi}) \approx 0 \tag{7.15}$$

for f_0 not near 0 or 1/2. Thus, the left-hand side of (7.14) when divided by N and set equal to zero will produce an approximate MLE, which satisfies

$$\sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n + \hat{\phi}) = 0. \tag{7.16}$$

Upon expanding this we have

$$\sum_{n=0}^{N-1} x[n]\sin 2\pi f_0 n \cos\hat{\phi} = -\sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n \sin\hat{\phi}$$

or finally the MLE of phase is given approximately as

$$\hat{\phi} = -\arctan\frac{\displaystyle\sum_{n=0}^{N-1} x[n]\sin 2\pi f_0 n}{\displaystyle\sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n}. \tag{7.17}$$

It is interesting to note that the MLE is a function of the sufficient statistics. In hindsight, this should not be surprising if we keep in mind the Neyman-Fisher factorization theorem. In this example there are two sufficient statistics, effecting a factorization as

$$p(\mathbf{x};\phi) = g(T_1(\mathbf{x}), T_2(\mathbf{x}), \phi)h(\mathbf{x}).$$

Clearly, maximizing $p(\mathbf{x};\phi)$ is equivalent to maximizing g ($h(\mathbf{x})$ can always be chosen so that $h(\mathbf{x}) > 0$), and thus $\hat{\phi}$ must be a function of $T_1(\mathbf{x})$ and $T_2(\mathbf{x})$.

According to Theorem 7.1, the asymptotic PDF of the phase estimator is

$$\hat{\phi} \overset{a}{\sim} \mathcal{N}(\phi, I^{-1}(\phi)). \tag{7.18}$$

TABLE 7.2 Theoretical Asymptotic and Actual Mean and Variance for Estimator in Example 7.6

Data Record Length, N	Mean, $E(\hat{\phi})$	$N\times$Variance, $N\,\text{var}(\hat{\phi})$
20	0.732	0.0978
40	0.746	0.108
60	0.774	0.110
80	0.789	0.0990
Theoretical asymptotic value	$\phi = 0.785$	$1/\eta = 0.1$

From Example 3.4

$$I(\phi) = \frac{NA^2}{2\sigma^2}$$

so that the asymptotic variance is

$$\text{var}(\hat{\phi}) = \frac{1}{N\frac{A^2}{2\sigma^2}} = \frac{1}{N\eta} \tag{7.19}$$

where $\eta = (A^2/2)/\sigma^2$ is the SNR. To determine the data record length for the asymptotic mean and variance to apply we performed a computer simulation using $A = 1, f_0 = 0.08, \phi = \pi/4$, and $\sigma^2 = 0.05$. The results are listed in Table 7.2. It is seen that the asymptotic mean and normalized variance are attained for $N = 80$. For shorter data records the estimator is considerably biased. Part of the bias is due to the assumption made in (7.15). The MLE given by (7.17) is actually valid only for large N. To find the exact MLE we would have to minimize J as given by (7.13) by evaluating it for all ϕ. This could be done by using a grid search to find the minimum. Also, observe from the table that the variance for $N = 20$ is below the CRLB. This is possible due to the bias of the estimator, thereby invalidating the CRLB which assumes an unbiased estimator.

Next we fixed the data record length at $N = 80$ and varied the SNR. Plots of the mean and variance versus the SNR, as well as the asymptotic values, are shown in Figures 7.3 and 7.4. As shown in Figure 7.3, the estimator attains the asymptotic mean above about -10 dB. In Figure 7.4 we have plotted $10\log_{10}\text{var}(\hat{\phi})$. This has the desirable effect of causing the CRLB to be a straight line when plotted versus the SNR in dB. In particular, from (7.19) the asymptotic variance or CRLB is

$$\begin{aligned}10\log_{10}\text{var}(\hat{\phi}) &= 10\log_{10}\frac{1}{N\eta}\\ &= -10\log_{10}N - 10\log_{10}\eta.\end{aligned}$$

Examining the results, we see a peculiar trend. For low SNRs the variance is higher than the CRLB. Only at higher SNRs is the CRLB attained. Hence, the required data record length for the asymptotic results to apply also depends on the SNR. To

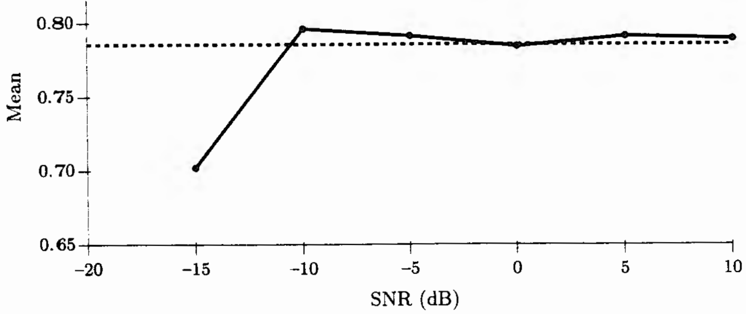


Figure 7.3 Actual vs. asymptotic mean for phase estimator

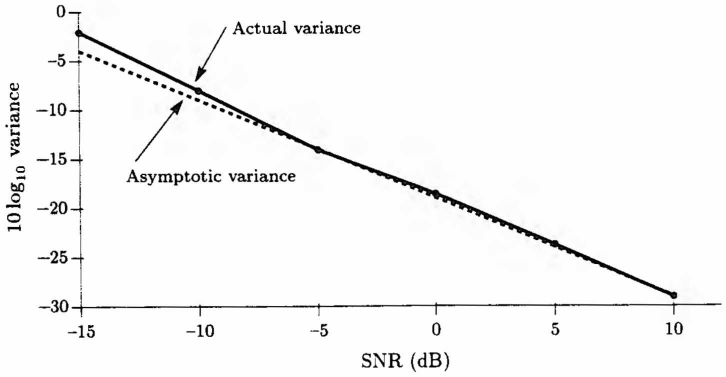
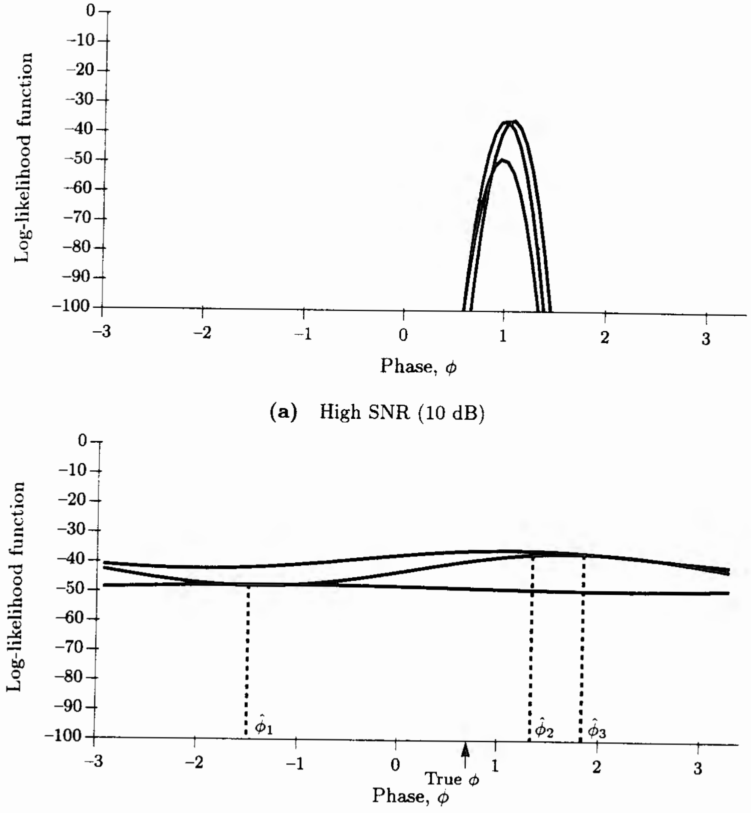


Figure 7.4 Actual vs. asymptotic variance for phase estimator

understand why this occurs we plot typical realizations of the log-likelihood function for different SNRs. As seen in Figure 7.5 for a high SNR, the maximum is relatively stable from realization to realization, and hence the MLE exhibits a low variance. For lower SNRs, however, the effect of the increased noise is to cause other peaks to occur. Occasionally these peaks are larger than the peak near the true value, causing a large estimation error and ultimately a larger variance. These large error estimates are said to be *outliers* and cause the *threshold* effect seen in Figures 7.3 and 7.4. Nonlinear estimators nearly always exhibit this effect. ◇



(a) High SNR (10 dB)

(b) Low SNR (-15 dB)

Figure 7.5 Typical realizations of log-likelihood function for phase

In summary, the asymptotic PDF of the MLE is valid for large enough data records. For signal in noise problems the CRLB may be attained even for short data records if the SNR is high enough. To see why this is so the phase estimator can be written from (7.17) as

$$\hat{\phi} = -\arctan \frac{\sum_{n=0}^{N-1} [A\cos(2\pi f_0 n + \phi) + w[n]] \sin 2\pi f_0 n}{\sum_{n=0}^{N-1} [A\cos(2\pi f_0 n + \phi) + w[n]] \cos 2\pi f_0 n}$$

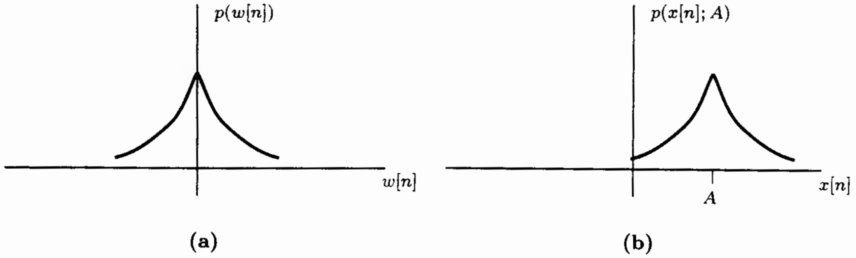


Figure 7.6 Non-Gaussian PDF for Example 7.7

$$\approx \quad -\arctan \frac{-\frac{NA}{2}\sin\phi + \sum_{n=0}^{N-1} w[n]\sin 2\pi f_0 n}{\frac{NA}{2}\cos\phi + \sum_{n=0}^{N-1} w[n]\cos 2\pi f_0 n}$$

where we have used the same type of approximation as in (7.15) and some standard trigonometric identities. Simplifying, we have

$$\hat{\phi} \approx \arctan \frac{\sin\phi - \frac{2}{NA}\sum_{n=0}^{N-1} w[n]\sin 2\pi f_0 n}{\cos\phi + \frac{2}{NA}\sum_{n=0}^{N-1} w[n]\cos 2\pi f_0 n}. \tag{7.20}$$

If the data record is large and/or the sinusoidal power is large, the noise terms will be small. *It is this condition, that the estimation error is small, that allows the MLE to attain its asymptotic distribution.* See also Problem 7.15 for a further discussion of this point.

In some cases the asymptotic distribution does not hold, no matter how large the data record and/or the SNR becomes. This tends to occur when the estimation error cannot be reduced due to a lack of averaging in the estimator. An example follows.

Example 7.7 - DC Level in Nonindependent Non-Gaussian Noise

Consider the observations

$$x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1$$

where each sample of $w[n]$ has the PDF $p(w[n])$ as shown in Figure 7.6a. The PDF is symmetric about $w[n] = 0$ and has a maximum at $w[n] = 0$. Furthermore, we assume all the noise samples are equal or $w[0] = w[1] = \cdots = w[N-1]$. In estimating A we need to consider only a single observation since all observations are identical. Utilizing only $x[0]$, we first note that the PDF of $x[0]$ is a shifted version of $p(w[n])$, where the shift is A as shown in Figure 7.6b. This is because $p_{x[0]}(x[0];A) = p_{w[0]}(x[0]-A)$. The MLE of A is the value that maximizes $p_{w[0]}(x[0]-A)$, which because the PDF of $w[0]$ has a maximum at $w[0] = 0$ becomes

$$\hat{A} = x[0].$$

This estimator has the mean

$$E(\hat{A}) = E(x[0]) = A$$

since the noise PDF is symmetric about $w[0] = 0$. The variance of $\hat{A}$ is the same as the variance of $x[0]$ or of $w[0]$. Hence,

$$\mathrm{var}(\hat{A}) = \int_{-\infty}^{\infty} u^2 p_{w[0]}(u)\,du$$

while the CRLB, is from Problem 3.2,

$$\mathrm{var}(\hat{A}) \geq \left[\int_{-\infty}^{\infty} \frac{\left(\frac{dp_{w[0]}(u)}{du}\right)^2}{p_{w[0]}(u)} du\right]^{-1}$$

and the two are not in general equal (see Problem 7.16). In this example, then, the estimation error does not decrease as the data record length increases but remains the same. Furthermore, the PDF of $\hat{A} = x[0]$ as shown in Figure 7.6b is a shifted version of $p(w[n])$, clearly not Gaussian. Finally, $\hat{A}$ is not even consistent as $N \to \infty$. ◇

7.6 MLE for Transformed Parameters

In many instances we wish to estimate a *function* of θ, the parameter characterizing the PDF. For example, we may not be interested in the value of a DC level A in WGN but only in the power A^2. In such a situation the MLE of A^2 is easily found from the MLE of A. Some examples illustrate how this is done.

Example 7.8 - Transformed DC Level in WGN

Consider the data

$$x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $w[n]$ is WGN with variance σ^2. We wish to find the MLE of $\alpha = \exp(A)$. The PDF is given as

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \qquad -\infty < A < \infty \tag{7.21}$$

and is parameterized by the parameter $\theta = A$. However, since α is a one-to-one transformation of A, we can equivalently parameterize the PDF as

$$p_T(\mathbf{x};\alpha) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\ln\alpha)^2\right] \quad \alpha > 0 \qquad (7.22)$$

where the subscript T indicates that the PDF is parameterized according to the *transformed* parameter. Clearly, $p_T(\mathbf{x};\alpha)$ is the PDF of the data set

$$x[n] = \ln\alpha + w[n] \qquad n = 0, 1, \ldots, N-1$$

and (7.21) and (7.22) are entirely equivalent. The MLE of α is found by maximizing (7.22) over α. Setting the derivative of $p_T(\mathbf{x};\alpha)$ with respect to α equal to zero yields

$$\sum_{n=0}^{N-1}(x[n]-\ln\hat{\alpha})\frac{1}{\hat{\alpha}} = 0$$

or

$$\hat{\alpha} = \exp(\bar{x}).$$

But $\bar{x}$ is just the MLE of A, so that

$$\hat{\alpha} = \exp(\hat{A}) = \exp(\hat{\theta}).$$

The MLE of the transformed parameter is found by substituting the MLE of the original parameter into the transformation. This property of the MLE is termed the ***invariance property.***

Example 7.9 - Transformed DC Level in WGN (Another Example)

Now consider the transformation $\alpha = A^2$ for the data set in the previous example. If we try to repeat the steps, we soon encounter a problem. Attempting to parameterize $p(\mathbf{x};A)$ with respect to α, we find that

$$A = \pm\sqrt{\alpha}$$

since the transformation is not one-to-one. If we choose $A = \sqrt{\alpha}$, then some of the possible PDFs of (7.21) will be missing. We actually require two sets of PDFs

$$\begin{aligned} p_{T_1}(\mathbf{x};\alpha) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\sqrt{\alpha})^2\right] \quad \alpha \ge 0 \\ p_{T_2}(\mathbf{x};\alpha) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]+\sqrt{\alpha})^2\right] \quad \alpha > 0 \qquad (7.23) \end{aligned}$$

to characterize all possible PDFs. It is possible to find the MLE of α as the value of α that yields the maximum of $p_{T_1}(\mathbf{x};\alpha)$ and $p_{T_2}(\mathbf{x};\alpha)$ or

$$\hat{\alpha} = \arg\max_{\alpha}\{p_{T_1}(\mathbf{x};\alpha), p_{T_2}(\mathbf{x};\alpha)\}. \qquad (7.24)$$

Alternatively, we can find the maximum in two steps as

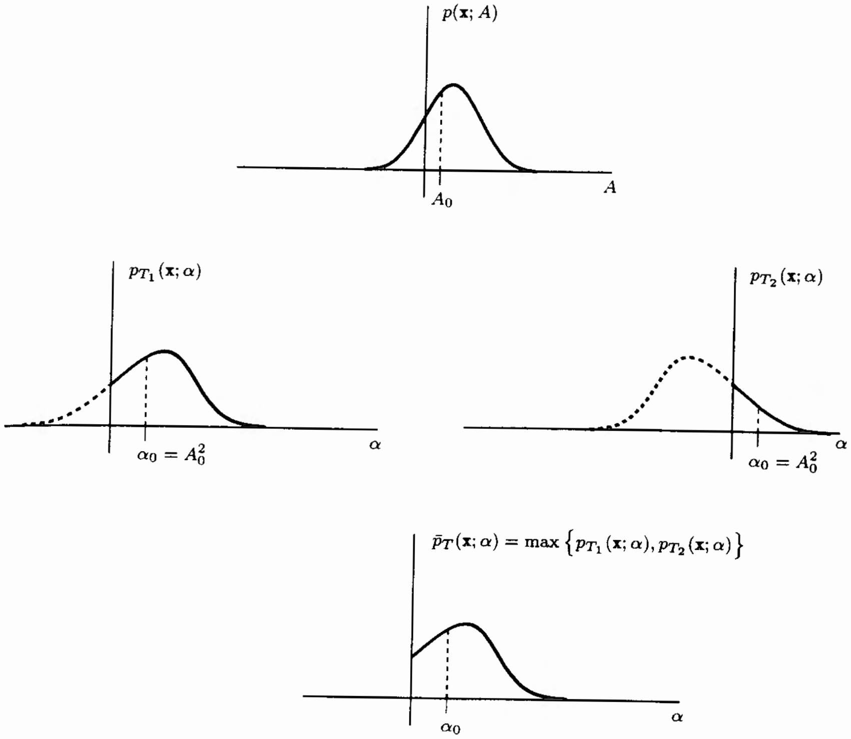


Figure 7.7 Construction of modified likelihood function

1. For a given value of α, say α_0, determine whether $p_{T_1}(\mathbf{x};\alpha)$ or $p_{T_2}(\mathbf{x};\alpha)$ is larger. If, for example,

$$p_{T_1}(\mathbf{x};\alpha_0) > p_{T_2}(\mathbf{x};\alpha_0),$$

then denote the value of $p_{T_1}(\mathbf{x};\alpha_0)$ as $\bar{p}_T(\mathbf{x};\alpha_0)$. Repeat for all $\alpha > 0$ to form $\bar{p}_T(\mathbf{x};\alpha)$. (Note that $\bar{p}_T(\mathbf{x};\alpha=0) = p(\mathbf{x};A=0)$.)

2. The MLE is given as the α that maximizes $\bar{p}_T(\mathbf{x};\alpha)$ over $\alpha \ge 0$.

This procedure is illustrated in Figure 7.7. The function $\bar{p}_T(\mathbf{x};\alpha)$ can be thought of as a *modified likelihood function*, having been derived from the original likelihood function by transforming the value of A that yields the maximum value for a given α. In this example, for each α the possible values of A are $\pm\sqrt{\alpha}$. Now, from (7.24) the MLE $\hat{\alpha}$ is

$$\hat{\alpha} \;=\; \arg\max_{\alpha\ge 0}\{p(\mathbf{x};\sqrt{\alpha}), p(\mathbf{x};-\sqrt{\alpha})\}$$

$$
\begin{aligned}
&= \left[\arg\max_{\sqrt{\alpha}\geq 0}\left\{p(\mathbf{x};\sqrt{\alpha}),p(\mathbf{x};-\sqrt{\alpha})\right\}\right]^2\\
&= \left[\arg\max_{-\infty<A<\infty} p(\mathbf{x};A)\right]^2\\
&= \hat{A}^2\\
&= \bar{x}^2
\end{aligned}
$$

so that again the invariance property holds. The understanding is that $\hat{\alpha}$ maximizes the *modified* likelihood function $\bar{p}_T(\mathbf{x};\alpha)$ since the standard likelihood function with α as a parameter cannot be defined. ◇

We summarize the preceding discussion in the following theorem.

Theorem 7.2 (Invariance Property of the MLE) *The MLE of the parameter $\alpha = g(\theta)$, where the PDF $p(\mathbf{x};\theta)$ is parameterized by θ, is given by*

$$\hat{\alpha} = g(\hat{\theta})$$

where $\hat{\theta}$ is the MLE of θ. The MLE of $\hat{\theta}$ is obtained by maximizing $p(\mathbf{x};\theta)$. If g is not a one-to-one function, then $\hat{\alpha}$ maximizes the modified likelihood function $\bar{p}_T(\mathbf{x};\alpha)$, defined as

$$\bar{p}_T(\mathbf{x};\alpha) = \max_{\{\theta:\alpha=g(\theta)\}} p(\mathbf{x};\theta).$$

We complete our discussion by giving another example.

Example 7.10 - Power of WGN in dB

We observe N samples of WGN with variance σ^2 whose power in dB is to be estimated. To do so we first find the MLE of σ^2. Then, we use the invariance principle to find the power P in dB, which is defined as

$$P = 10\log_{10}\sigma^2.$$

The PDF is given by

$$p(\mathbf{x};\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right].$$

Differentiating the log-likelihood function produces

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{x};\sigma^2)}{\partial\sigma^2} &= \frac{\partial}{\partial\sigma^2}\left[-\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]\\
&= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=0}^{N-1}x^2[n]
\end{aligned}
$$

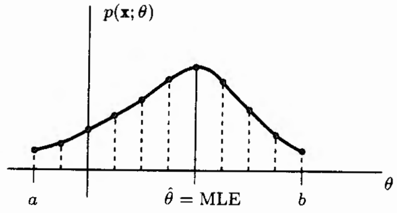


Figure 7.8 Grid search for MLE

and upon setting it equal to zero yields the MLE

$$\hat{\sigma^2} = \frac{1}{N}\sum_{n=0}^{N-1}x^2[n].$$

The MLE of the power in dB readily follows as

$$
\begin{aligned}
\hat{P} &= 10\log_{10}\hat{\sigma^2}\\
&= 10\log_{10}\frac{1}{N}\sum_{n=0}^{N-1}x^2[n].
\end{aligned}
$$

◇

7.7 Numerical Determination of the MLE

A distinct advantage of the MLE is that we can always find it for a given data set *numerically*. This is because the MLE is determined as the maximum of a known function, namely, the likelihood function. If, for example, the allowable values of θ lie in the interval $[a,b]$, then we need only maximize $p(\mathbf{x};\theta)$ over that interval. The "safest" way to do this is to perform a grid search over the $[a,b]$ interval as shown in Figure 7.8. As long as the spacing between θ values is small enough, we are guaranteed to find the MLE for the given set of data. Of course, for a new set of data we will have to repeat the search since the likelihood function will undoubtedly change. If, however, the range of θ is not confined to a finite interval, as in estimating the variance of a noise process for which $\sigma^2 > 0$, then a grid search may not be computationally feasible. In such a case, we are forced to resort to iterative maximization procedures. Some typical ones are the Newton-Raphson method, the scoring approach, and the expectation-maximization algorithm. In general, these methods will produce the MLE if the initial guess is close to the true maximum. If not, convergence may not be attained, or only convergence to a local maximum. The difficulty with the use of these iterative methods is that in general we do not know beforehand if they will converge and, even if convergence is attained, whether the value produced is the MLE. An important distinction of our estimation problem which sets it apart from other maximization problems is that the function to be maximized is *not known a priori*. The likelihood function changes for each data set, requiring the maximization of a *random function*. Nevertheless, these methods can at times produce good results. We now describe some of the more common ones. The interested reader is referred to [Bard 1974] for a more complete description of methods for nonlinear optimization as applied to estimation problems.

As a means of comparison, we will apply the methods to the following example.

Example 7.11 - Exponential in WGN

Consider the data

$$x[n] = r^n + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $w[n]$ is WGN with variance σ^2. The parameter r, the exponential factor, is to be estimated. Allowable values are $r > 0$. The MLE of r is the value that maximizes the likelihood function

$$p(\mathbf{x}; r) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - r^n)^2\right]$$

or, equivalently, the value that minimizes

$$J(r) = \sum_{n=0}^{N-1}(x[n] - r^n)^2 .$$

Differentiating $J(r)$ and setting it equal to zero produces

$$\sum_{n=0}^{N-1}(x[n] - r^n)\, nr^{n-1} = 0. \qquad (7.25)$$

This is a nonlinear equation in r and cannot be solved directly. We will now consider the iterative methods of Newton-Raphson and scoring. ◇

The iterative methods attempt to maximize the log-likelihood function by finding a zero of the derivative function. To do so the derivative is taken and set equal to zero, yielding

$$\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} = 0. \qquad (7.26)$$

Then, the methods attempt to solve this equation iteratively. Let

$$g(\theta) = \frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta}$$

and assume that we have an initial guess for the solution to (7.26). Call this guess θ_0. Then, if $g(\theta)$ is approximately linear near θ_0, we can approximate it by

$$g(\theta) \approx g(\theta_0) + \left.\frac{dg(\theta)}{d\theta}\right|_{\theta=\theta_0}(\theta - \theta_0) \qquad (7.27)$$

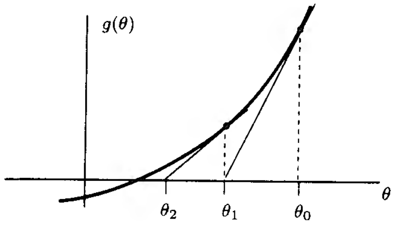


Figure 7.9 Newton-Raphson method for finding zero of function

as shown in Figure 7.9. Next, we use (7.27) to solve for the zero θ_1, so that upon setting $g(\theta_1)$ equal to zero and solving for θ_1 we have

$$\theta_1 = \theta_0 - \frac{g(\theta_0)}{\left.\dfrac{dg(\theta)}{d\theta}\right|_{\theta=\theta_0}}.$$

Again we linearize g but use the new guess, θ_1, as our point of linearization and repeat the previous procedure to find the new zero. As shown in Figure 7.9, the sequence of guesses will converge to the true zero of $g(\theta)$. In general, the Newton-Raphson iteration finds the new guess, θ_{k+1}, based on the previous one, θ_k, using

$$\theta_{k+1} = \theta_k - \frac{g(\theta_k)}{\left.\dfrac{dg(\theta)}{d\theta}\right|_{\theta=\theta_k}}. \qquad (7.28)$$

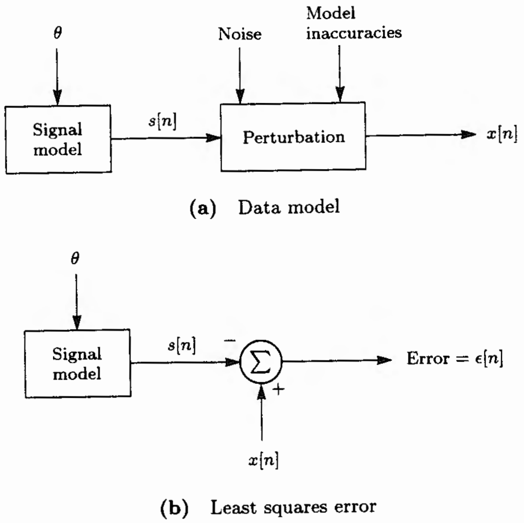
Note that at convergence $\theta_{k+1} = \theta_k$, and from (7.28) $g(\theta_k) = 0$, as desired. Since $g(\theta)$ is the derivative of the log-likelihood function, we find the MLE as

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2}\right]^{-1} \left.\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta}\right|_{\theta=\theta_k} . \qquad (7.29)$$

Several points need to be raised concerning the Newton-Raphson iterative procedure.

1. The iteration may not converge. This will be particularly evident when the second derivative of the log-likelihood function is small. In this case it is seen from (7.29) that the correction term may fluctuate wildly from iteration to iteration.

2. Even if the iteration converges, the point found may not be the global maximum but possibly only a local maximum or even a local minimum. Hence, to avoid these possibilities it is best to use several starting points and at convergence choose the one that yields the maximum. Generally, if the initial point is close to the global maximum, the iteration will converge to it. *The importance of a good initial guess cannot be overemphasized.* An illustration is given in Problem 7.18.



(a) Data model

(b) Least squares error

Figure 8.1 Least squares approach

available, then a sequential approach can be employed. It determines the least squares estimator based on the estimate at the previous time and the new data. The equations (8.46)-(8.48) summarize the calculations required. At times the parameter vector is constrained, as in (8.50). In such a case the constrained least squares estimator is given by (8.52). Nonlinear least squares is discussed in Section 8.9. Some methods for converting the problem to a linear one are described, followed by iterative minimization approaches if this is not possible. The two methods that are generally used are the Newton-Raphson iteration of (8.61) and the Gauss-Newton iteration of (8.62).

8.3 The Least Squares Approach

Our focus in determining a good estimator has been to find one that was unbiased and had minimum variance. In choosing the variance as our measure of goodness we implicitly sought to minimize the discrepancy (on the average) between our estimate and the true parameter value. In the least squares (LS) approach we attempt to minimize the squared difference between the given data $x[n]$ and the assumed signal or noiseless data. This is illustrated in Figure 8.1. The signal is generated by some model which in turn depends upon our unknown parameter θ. The signal $s[n]$ is purely deterministic. Due to observation noise or model inaccuracies we observe a perturbed version of $s[n]$, which we denote by $x[n]$. The least squares estimator (LSE) of θ chooses the value that

makes $s[n]$ closest to the observed data $x[n]$. Closeness is measured by the LS error criterion

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - s[n])^2 \tag{8.1}$$

where the observation interval is assumed to be $n = 0, 1, \ldots, N-1$, and the dependence of J on θ is via $s[n]$. The value of θ that minimizes $J(\theta)$ is the LSE. *Note that no probabilistic assumptions have been made about the data $x[n]$.* The method is equally valid for Gaussian as well as non-Gaussian noise. Of course, the *performance* of the LSE will undoubtedly depend upon the properties of the corrupting noise as well as any modeling errors. LSEs are usually applied in situations where a precise statistical characterization of the data is unknown or where an optimal estimator cannot be found or may be too complicated to apply in practice.

Example 8.1 - DC Level Signal

Assume that the signal model in Figure 8.1 is $s[n] = A$ and we observe $x[n]$ for $n = 0, 1, \ldots, N-1$. Then, according to the LS approach, we can estimate A by minimizing (8.1) or

$$J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2.$$

Differentiating with respect to A and setting the result equal to zero produces

$$\begin{aligned} \hat{A} &= \frac{1}{N} \sum_{n=0}^{N-1} x[n] \\ &= \bar{x} \end{aligned}$$

or the sample mean estimator. Our familiar estimator, however, cannot be claimed to be optimal in the MVU sense but only in that it minimizes the LS error. We know, however, from our previous discussions that if $x[n] = A + w[n]$, where $w[n]$ is zero mean WGN, then the LSE will also be the MVU estimator, but otherwise not. To underscore the potential difficulties, consider what would happen if the noise were not zero mean. Then, the sample mean estimator would actually be an estimator of $A + E(w[n])$ since $w[n]$ could be written as

$$w[n] = E(w[n]) + w'[n]$$

where $w'[n]$ is zero mean noise. The data are more appropriately described by

$$x[n] = A + E(w[n]) + w'[n].$$

It should be clear that in using this approach, it must be assumed that the observed data are composed of a deterministic signal and *zero mean* noise. If this is the case, the error $\epsilon[n] = x[n] - s[n]$ will tend to be zero on the average for the correct choice of the signal parameters. The minimization of (8.1) is then a reasonable approach. The reader might also consider what would happen if the assumed DC level signal model were incorrect, as for instance if $x[n] = A+Bn+w[n]$ described the data. This modeling error would also cause the LSE to be biased. ◇

Example 8.2 - Sinusoidal Frequency Estimation

Consider the signal model

$$s[n] = \cos 2\pi f_0 n$$

in which the frequency f_0 is to be estimated. The LSE is found by minimizing

$$J(f_0) = \sum_{n=0}^{N-1}(x[n] - \cos 2\pi f_0 n)^2.$$

In contrast to the DC level signal for which the minimum is easily found, here the LS error is highly nonlinear in f_0. The minimization cannot be done in closed form. Since the error criterion is a quadratic function of the signal, a signal that is *linear* in the unknown parameter yields a quadratic function for J, as in the previous example. The minimization is then easily carried out. A signal model that is *linear in the unknown parameter* is said to generate a *linear least squares* problem. Otherwise, as in this example, the problem is a *nonlinear least squares* problem. Nonlinear LS problems are solved via grid searches or iterative minimization methods as described in Section 8.9. It should be noted that the signal itself need not be linear but only in the unknown parameter, as the next example illustrates. ◇

Example 8.3 - Sinusoidal Amplitude Estimation

If the signal is $s[n] = A\cos 2\pi f_0 n$, where f_0 *is known* and A is to be estimated, then the LSE minimizes

$$J(A) = \sum_{n=0}^{N-1}(x[n] - A\cos 2\pi f_0 n)^2$$

over A. This is easily accomplished by differentiation since $J(A)$ is quadratic in A. This linear LS problem is clearly very desirable from a practical viewpoint. If, however, A were known and the frequency were to be estimated, the problem would be equivalent to that in Example 8.2, that is to say, it would be a nonlinear LS problem. A final possibility, in the vector parameter case, is that both A and f_0 might need to be estimated. Then, the error criterion

$$J(A, f_0) = \sum_{n=0}^{N-1}(x[n] - A\cos 2\pi f_0 n)^2$$

is quadratic in A but nonquadratic in f_0. The net result is that J can be minimized in closed form with respect to A for a given f_0, reducing the minimization of J to one over f_0 only. This type of problem, in which the signal is linear in some parameters but nonlinear in others, is termed a *separable least squares* problem. In Section 8.9 we will discuss this further. ◇

8.4 Linear Least Squares

In applying the linear LS approach for a scalar parameter we must assume that

$$s[n] = \theta h[n] \tag{8.2}$$

where $h[n]$ is a known sequence. (The reader may want to refer back to Chapter 6 on the BLUE for a comparison to similar signal models.) The LS error criterion becomes

$$J(\theta) = \sum_{n=0}^{N-1}(x[n] - \theta h[n])^2. \tag{8.3}$$

A minimization is readily shown to produce the LSE

$$\hat{\theta} = \frac{\sum_{n=0}^{N-1} x[n]h[n]}{\sum_{n=0}^{N-1} h^2[n]}. \tag{8.4}$$

The minimum LS error, obtained by substituting (8.4) into (8.3), is

$$\begin{aligned} J_{\min} = J(\hat{\theta}) &= \sum_{n=0}^{N-1}(x[n] - \hat{\theta}h[n])(x[n] - \hat{\theta}h[n]) \\ &= \sum_{n=0}^{N-1} x[n](x[n] - \hat{\theta}h[n]) - \underbrace{\hat{\theta}\sum_{n=0}^{N-1} h[n](x[n] - \hat{\theta}h[n])}_{S} \\ &= \sum_{n=0}^{N-1} x^2[n] - \hat{\theta}\sum_{n=0}^{N-1} x[n]h[n]. \end{aligned} \tag{8.5}$$

The last step follows because the sum S is zero (substitute $\hat{\theta}$ to verify). Alternatively, by using (8.4) we can rewrite $J_{\min}$ as

$$J_{\min} = \sum_{n=0}^{N-1} x^2[n] - \frac{\left(\sum_{n=0}^{N-1} x[n]h[n]\right)^2}{\sum_{n=0}^{N-1} h^2[n]}. \tag{8.6}$$

The minimum LS error is the original energy of the data or $\sum_{n=0}^{N-1} x^2[n]$ less that due to the signal fitting. For Example 8.1 in which $\theta = A$ we have $h[n] = 1$, so that from (8.4) $\hat{A} = \bar{x}$ and from (8.5)

$$J_{\min} = \sum_{n=0}^{N-1} x^2[n] - N\bar{x}^2.$$

If the data were noiseless so that $x[n] = A$, then $J_{\min} = 0$ or we would have a perfect LS fit to the data. On the other hand, if $x[n] = A + w[n]$, where $E(w^2[n]) \gg A^2$, then $\sum_{n=0}^{N-1} x^2[n]/N \gg \bar{x}^2$. The minimum LS error would then be

$$J_{\min} \approx \sum_{n=0}^{N-1} x^2[n]$$

or not much different than the original error. It can be shown (see Problem 8.2) that the minimum LS error is always between these two extremes or

$$0 \le J_{\min} \le \sum_{n=0}^{N-1} x^2[n]. \tag{8.7}$$

The extension of these results to a vector parameter $\boldsymbol{\theta}$ of dimension $p \times 1$ is straightforward and of great practical utility. For the signal $\mathbf{s} = [s[0]\, s[1] \ldots s[N-1]]^T$ to be linear in the unknown parameters, we assume, using matrix notation,

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta} \tag{8.8}$$

where $\mathbf{H}$ is a known $N \times p$ matrix $(N > p)$ of full rank p. The matrix $\mathbf{H}$ is referred to as the *observation matrix*. This is, of course, the linear model, albeit without the usual noise PDF assumption. Many examples of signals satisfying this model can be found in Chapter 4. The LSE is found by minimizing

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum_{n=0}^{N-1} (x[n] - s[n])^2 \\ &= (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}). \end{aligned} \tag{8.9}$$

This is easily accomplished (since J is a quadratic function of $\boldsymbol{\theta}$) by using (4.3). Since

$$\begin{aligned} J(\boldsymbol{\theta}) &= \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{H}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{x} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} \\ &= \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} \end{aligned}$$

(note that $\mathbf{x}^T\mathbf{H}\boldsymbol{\theta}$ is a scalar), the gradient is

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}.$$

Setting the gradient equal to zero yields the LSE

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}. \tag{8.10}$$

The equations $\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} = \mathbf{H}^T\mathbf{x}$ to be solved for $\hat{\boldsymbol{\theta}}$ are termed the *normal equations.* The assumed full rank of $\mathbf{H}$ guarantees the invertibility of $\mathbf{H}^T\mathbf{H}$. See also Problem 8.4 for another derivation. Somewhat surprisingly, we obtain an estimator that has the *identical functional form* as the efficient estimator for the linear model as well as the BLUE. That $\hat{\boldsymbol{\theta}}$ as given by (8.10) is not the *identical estimator* stems from the assumptions made about the data. For it to be the BLUE would require $E(\mathbf{x}) = \mathbf{H}\boldsymbol{\theta}$ and $\mathbf{C}_x = \sigma^2\mathbf{I}$ (see Chapter 6), and to be efficient would in addition to these properties require $\mathbf{x}$ to be Gaussian (see Chapter 4). As a side issue, if these assumptions hold, we can easily determine the statistical properties of the LSE (see Problem 8.6), having been given in Chapters 4 and 6. Otherwise, this may be quite difficult. The minimum LS error is found from (8.9) and (8.10) as

$$\begin{aligned} J_{\min} &= J(\hat{\boldsymbol{\theta}}) \\ &= (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})^T(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) \\ &= \left(\mathbf{x} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}\right)^T \left(\mathbf{x} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}\right) \\ &= \mathbf{x}^T\left(\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\right)\left(\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\right)\mathbf{x} \\ &= \mathbf{x}^T\left(\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\right)\mathbf{x}. \end{aligned} \tag{8.11}$$

The last step results from the fact that $\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ is an idempotent matrix or it has the property $\mathbf{A}^2 = \mathbf{A}$. Other forms for $J_{\min}$ are

$$J_{\min} = \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \tag{8.12}$$

$$\phantom{J_{\min}} = \mathbf{x}^T(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}). \tag{8.13}$$

An extension of the linear LS problem is to *weighted* LS. Instead of minimizing (8.9), we include an $N \times N$ positive definite (and by definition therefore symmetric) weighting matrix $\mathbf{W}$, so that

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T\mathbf{W}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}). \tag{8.14}$$

If, for instance, $\mathbf{W}$ is diagonal with diagonal elements $[\mathbf{W}]_{ii} = w_i > 0$, then the LS error for Example 8.1 will be

$$J(A) = \sum_{n=0}^{N-1} w_n(x[n] - A)^2.$$

The rationale for introducing weighting factors into the error criterion is to emphasize the contributions of those data samples that are deemed to be more reliable. Again, considering Example 8.1, if $x[n] = A + w[n]$, where $w[n]$ is zero mean uncorrelated noise with variance σ_n^2, then it is reasonable to choose $w_n = 1/\sigma_n^2$. This choice will result in the estimator (see Problem 8.8)

$$\hat{A} = \frac{\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2}}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}}. \tag{8.15}$$

This familiar estimator is of course the BLUE since the $w[n]$'s are uncorrelated so that $\mathbf{W} = \mathbf{C}^{-1}$ (see Example 6.2).

The general form of the weighted LSE is readily shown to be

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{x} \tag{8.16}$$

and its minimum LS error is

$$J_{\min} = \mathbf{x}^T\left(\mathbf{W} - \mathbf{W}\mathbf{H}(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\right)\mathbf{x} \tag{8.17}$$

(see Problem 8.9).

8.5 Geometrical Interpretations

We now reexamine the linear LS approach from a geometrical perspective. This has the advantage of more clearly revealing the essence of the approach and leads to additional useful properties and insights into the estimator. Recall the general signal model $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$. If we denote the columns of $\mathbf{H}$ by $\mathbf{h}_i$, we have

$$\begin{aligned}\mathbf{s} &= \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \dots & \mathbf{h}_p \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \\ &= \sum_{i=1}^{p} \theta_i \mathbf{h}_i \end{aligned}$$

so that the signal model is seen to be a linear combination of the "signal" vectors $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_p\}$.

Example 8.4 - Fourier Analysis

Referring to Example 4.2 (with $M = 1$), we suppose the signal model to be

$$s[n] = a\cos 2\pi f_0 n + b \sin 2\pi f_0 n \qquad n = 0, 1, \dots, N-1$$

where f_0 is a known frequency and $\boldsymbol{\theta} = [a\, b]^T$ is to be estimated. Then, in vector form we have

$$\begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \cos 2\pi f_0 & \sin 2\pi f_0 \\ \vdots & \vdots \\ \cos 2\pi f_0(N-1) & \sin 2\pi f_0(N-1) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}. \tag{8.18}$$

It is seen that the columns of $\mathbf{H}$ are composed of the samples of the cosinusoidal and sinusoidal sequences. Alternatively, since

$$\begin{aligned}\mathbf{h}_1 &= \begin{bmatrix} 1 & \cos 2\pi f_0 & \dots & \cos 2\pi f_0(N-1) \end{bmatrix}^T \\ \mathbf{h}_2 &= \begin{bmatrix} 0 & \sin 2\pi f_0 & \dots & \sin 2\pi f_0(N-1) \end{bmatrix}^T, \end{aligned}$$

we have

$$\mathbf{s} = a\mathbf{h}_1 + b\mathbf{h}_2.$$

◇

The LS error was defined to be

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}).$$

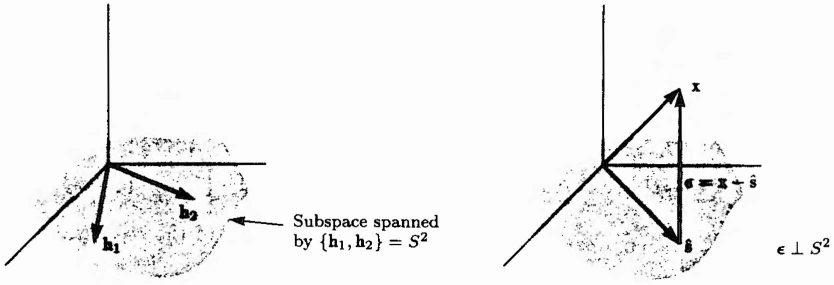
If we further define the Euclidean length of an $N \times 1$ vector $\boldsymbol{\xi} = [\xi_1\, \xi_2 \dots \xi_N]^T$ as

$$||\boldsymbol{\xi}|| = \sqrt{\sum_{i=1}^{N} \xi_i^2} = \sqrt{\boldsymbol{\xi}^T\boldsymbol{\xi}},$$

then the LS error can be also written as

$$\begin{aligned} J(\boldsymbol{\theta}) &= ||\mathbf{x} - \mathbf{H}\boldsymbol{\theta}||^2 \\ &= ||\mathbf{x} - \sum_{i=1}^{p} \theta_i \mathbf{h}_i||^2. \end{aligned} \tag{8.19}$$

We now see that the linear LS approach attempts to minimize the square of the distance from the data vector $\mathbf{x}$ to a signal vector $\sum_{i=1}^{p} \theta_i \mathbf{h}_i$, which must be a linear combination of the columns of $\mathbf{H}$. The data vector can lie anywhere in an N-dimensional space, termed R^N, while all possible signal vectors, being linear combinations of $p < N$ vectors, must lie in a p-dimensional subspace of R^N, termed S^p. (The full rank of $\mathbf{H}$ assumption assures us that the columns are linearly independent and hence the subspace spanned is truly p-dimensional.) For $N = 3$ and $p = 2$ we illustrate this in Figure 8.2. Note that all possible choices of θ_1, θ_2 (where we assume $-\infty < \theta_1 < \infty$ and $-\infty < \theta_2 < \infty$) produce signal vectors constrained to lie in the subspace S^2 and that in general $\mathbf{x}$ does not lie in the subspace. It should be intuitively clear that the vector $\hat{\mathbf{s}}$ that lies in S^2 and


$\mathbf{h}_1$
$\mathbf{h}_2$
Subspace spanned by $\{\mathbf{h}_1, \mathbf{h}_2\} = S^2$
$\mathbf{x}$
$\boldsymbol{\epsilon} = \mathbf{x} - \hat{\mathbf{s}}$
$\hat{\mathbf{s}}$
$\epsilon \perp S^2$

(a) Signal subspace

(b) Orthogonal projection to determine signal estimate

Figure 8.2 Geometrical viewpoint of linear least squares in $\mathbf{R}^3$

that is closest to $\mathbf{x}$ in the Euclidean sense is the component of $\mathbf{x}$ in S^2. Alternatively, $\hat{\mathbf{s}}$ is the *orthogonal projection* of $\mathbf{x}$ onto S^2. This means that the error vector $\mathbf{x} - \hat{\mathbf{s}}$ must be orthogonal to all vectors in S^2. Two vectors in R^N are defined to be orthogonal if $\mathbf{x}^T\mathbf{y} = 0$. To actually determine $\hat{\mathbf{s}}$ for this example we use the orthogonality condition. This says that the error vector is orthogonal to the signal subspace or

$$(\mathbf{x} - \hat{\mathbf{s}}) \perp S^2$$

where $\perp$ denotes orthogonal (or perpendicular). For this to be true we must have

$$\begin{aligned} (\mathbf{x} - \hat{\mathbf{s}}) &\perp \mathbf{h}_1 \\ (\mathbf{x} - \hat{\mathbf{s}}) &\perp \mathbf{h}_2 \end{aligned}$$

since then the error vector will be orthogonal to any linear combination of $\mathbf{h}_1$ and $\mathbf{h}_2$. Using the definition of orthogonality, we have

$$\begin{aligned} (\mathbf{x} - \hat{\mathbf{s}})^T\mathbf{h}_1 &= 0 \\ (\mathbf{x} - \hat{\mathbf{s}})^T\mathbf{h}_2 &= 0. \end{aligned}$$

Letting $\hat{\mathbf{s}} = \theta_1\mathbf{h}_1 + \theta_2\mathbf{h}_2$, we have

$$\begin{aligned} (\mathbf{x} - \theta_1\mathbf{h}_1 - \theta_2\mathbf{h}_2)^T\mathbf{h}_1 &= 0 \\ (\mathbf{x} - \theta_1\mathbf{h}_1 - \theta_2\mathbf{h}_2)^T\mathbf{h}_2 &= 0. \end{aligned}$$

In matrix form this is

$$\begin{aligned} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T\mathbf{h}_1 &= 0 \\ (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T\mathbf{h}_2 &= 0. \end{aligned}$$

Combining the two equations yields

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \left[\begin{array}{cc} \mathbf{h}_1 & \mathbf{h}_2 \end{array} \right] = \mathbf{0}^T$$

or

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T\mathbf{H} = \mathbf{0}^T. \tag{8.20}$$

Finally, we have as our LSE

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}.$$

Note that if $\boldsymbol{\epsilon} = \mathbf{x} - \mathbf{H}\boldsymbol{\theta}$ denotes the error vector, then the LSE is found from (8.20) by invoking the condition

$$\boldsymbol{\epsilon}^T\mathbf{H} = \mathbf{0}^T. \tag{8.21}$$

The error vector must be orthogonal to the columns of $\mathbf{H}$. This is the well-known *orthogonality principle.* In effect, the error represents the part of $\mathbf{x}$ that cannot be described by the signal model. A similar orthogonality principle will arise in Chapter 12 in our study of estimation of random parameters.

Again referring to Figure 8.2b, the minimum LS error is $||\mathbf{x} - \hat{\mathbf{s}}||^2$ or

$$\begin{aligned} ||\mathbf{x} - \hat{\mathbf{s}}||^2 &= ||\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}||^2 \\ &= (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})^T(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}). \end{aligned}$$

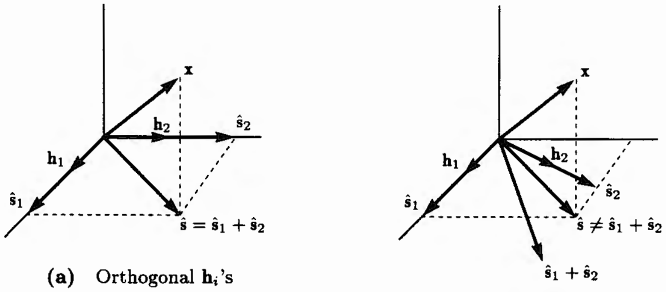
In evaluating this error we can make use of (8.21) (we have already done so for the scalar case in arriving at (8.5)). This produces

$$\begin{aligned} J_{\min} &= (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})^T(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) \\ &= \mathbf{x}^T(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}}^T\mathbf{H}^T\boldsymbol{\epsilon} \\ &= \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{H}\hat{\boldsymbol{\theta}} \\ &= \mathbf{x}^T\left(\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\right)\mathbf{x}. \end{aligned} \tag{8.22}$$

In summary, the LS approach can be interpreted as the problem of fitting or approximating a data vector $\mathbf{x}$ in R^N by another vector $\hat{\mathbf{s}}$, which is a linear combination of vectors $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_p\}$ that lie in a p-dimensional subspace of R^N. The problem is solved by choosing $\hat{\mathbf{s}}$ in the subspace to be the orthogonal projection of $\mathbf{x}$. Many of our intuitive notions about vector geometry may be used to our advantage once this connection is made. We now discuss some of these consequences.

Referring to Figure 8.3a, if it had happened that $\mathbf{h}_1$ and $\mathbf{h}_2$ were orthogonal, then $\hat{\mathbf{s}}$ could have easily been found. This is because the component of $\hat{\mathbf{s}}$ along $\mathbf{h}_1$ or $\hat{\mathbf{s}}_1$ does not contain a component of $\hat{\mathbf{s}}$ along $\mathbf{h}_2$. If it did, then we would have the situation in Figure 8.3b. Making the orthogonality assumption and also assuming that $||\mathbf{h}_1|| = ||\mathbf{h}_2|| = 1$ (ortho*normal* vectors), we have

$$\begin{aligned} \hat{\mathbf{s}} &= \hat{\mathbf{s}}_1 + \hat{\mathbf{s}}_2 \\ &= (\mathbf{h}_1^T\mathbf{x})\mathbf{h}_1 + (\mathbf{h}_2^T\mathbf{x})\mathbf{h}_2 \end{aligned}$$


x
$\mathbf{h}_2$
$\hat{\mathbf{s}}_2$
$\mathbf{h}_1$
$\hat{\mathbf{s}}_1$
$\hat{\mathbf{s}} = \hat{\mathbf{s}}_1 + \hat{\mathbf{s}}_2$
x
$\mathbf{h}_1$
$\mathbf{h}_2$
$\hat{\mathbf{s}}_1$
$\hat{\mathbf{s}}_2$
$\hat{\mathbf{s}} \neq \hat{\mathbf{s}}_1 + \hat{\mathbf{s}}_2$
$\hat{\mathbf{s}}_1 + \hat{\mathbf{s}}_2$

(a) Orthogonal $\mathbf{h}_i$'s

(b) Nonorthogonal $\mathbf{h}_i$'s

Figure 8.3 Effect of nonorthogonal columns of observation matrix

where $\mathbf{h}_i^T\mathbf{x}$ is the length of the vector $\mathbf{x}$ along $\mathbf{h}_i$. In matrix notation this is

$$\begin{aligned}\hat{\mathbf{s}} &= \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 \end{bmatrix} \begin{bmatrix} \mathbf{h}_1^T\mathbf{x} \\ \mathbf{h}_2^T\mathbf{x} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 \end{bmatrix} \begin{bmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \end{bmatrix} \mathbf{x} \\ &= \mathbf{H}\mathbf{H}^T\mathbf{x}\end{aligned}$$

so that

$$\hat{\boldsymbol{\theta}} = \mathbf{H}^T\mathbf{x}.$$

This result is due to the orthonormal columns of $\mathbf{H}$. As a result, we have

$$(\mathbf{H}^T\mathbf{H})^{-1} = (\mathbf{I})^{-1} = \mathbf{I}$$

and therefore

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} = \mathbf{H}^T\mathbf{x}.$$

No inversion is necessary. An example follows.

Example 8.5 - Fourier Analysis (continued)

Continuing Example 8.4, if $f_0 = k/N$, where k is an integer taking on any of the values $k = 1, 2, \ldots, N/2 - 1$, it is easily shown (see (4.13)) that

$$\mathbf{h}_1^T\mathbf{h}_2 = \sum_{n=0}^{N-1} \cos\left(2\pi\frac{k}{N}n\right) \sin\left(2\pi\frac{k}{N}n\right) = 0$$

and also

$$\begin{aligned}\mathbf{h}_1^T\mathbf{h}_1 &= \frac{N}{2} \\ \mathbf{h}_2^T\mathbf{h}_2 &= \frac{N}{2}\end{aligned}$$

so that $\mathbf{h}_1$ and $\mathbf{h}_2$ are orthogonal but not orthonormal. Combining these results produces $\mathbf{H}^T\mathbf{H} = (N/2)\mathbf{I}$, and therefore,

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} \\ &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \\ &= \frac{2}{N}\mathbf{H}^T\mathbf{x} \\ &= \begin{bmatrix} \dfrac{2}{N}\displaystyle\sum_{n=0}^{N-1} x[n]\cos\left(2\pi\frac{k}{N}n\right) \\ \dfrac{2}{N}\displaystyle\sum_{n=0}^{N-1} x[n]\sin\left(2\pi\frac{k}{N}n\right) \end{bmatrix}.\end{aligned}$$

If we had instead defined the signal as

$$s[n] = a'\sqrt{\frac{2}{N}}\cos\left(2\pi\frac{k}{N}n\right) + b'\sqrt{\frac{2}{N}}\sin\left(2\pi\frac{k}{N}n\right)$$

then the columns of $\mathbf{H}$ would have been orthonormal. ◇

In general, the columns of $\mathbf{H}$ will not be orthogonal, so that the signal vector estimate is obtained as

$$\hat{\mathbf{s}} = \mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}.$$

The signal estimate is the orthogonal projection of $\mathbf{x}$ onto the p-dimensional subspace. The $N \times N$ matrix $\mathbf{P} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ is known as the *orthogonal projection matrix* or just the *projection matrix*. It has the properties

1. $\mathbf{P}^T = \mathbf{P}$, symmetric

2. $\mathbf{P}^2 = \mathbf{P}$, idempotent.

That the projection matrix must be symmetric is shown in Problem 8.11, that it must be idempotent follows from the observation that if $\mathbf{P}$ is applied to $\mathbf{Px}$, then the same vector must result since $\mathbf{Px}$ is already in the subspace. Additionally, the projection matrix must be singular (for independent columns of $\mathbf{H}$ it has rank p, as shown in Problem 8.12). If it were not, then $\mathbf{x}$ could be recovered from $\hat{\mathbf{s}}$, which is clearly impossible since many $\mathbf{x}$'s have the same projection, as shown in Figure 8.4.