From Equation 6.4–12 we infer that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{6.4–14}$$

From Equation 6.4–13 we infer that, using the result from Equation 6.4–12,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2. \tag{6.4–15}$$

6.5 LINEAR ESTIMATION OF VECTOR PARAMETERS

A great many measurement problems in the real world are described by the following model

$$y(t) = \int_T h(t, \tau)\theta(\tau)\, d\tau + n(t), \tag{6.5–1}$$

where $y(t)$ is the *observation* or *measurement*, T is the integration set, $\theta(\tau)$ is the unknown *parameter* function, and $h(t, \tau)$ is a function that is characteristic of the system and links the parameter function to the measurement but is itself independent of $\theta(\tau)$, and $n(t)$ is the inevitable error in the measurement due to noise. For computational purposes Equation 6.5–1 must be reduced to its discrete form

$$\mathbf{Y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{N}, \tag{6.5–2}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations, $\mathbf{H}$ is an $n \times k$ matrix ($n > k$), $\boldsymbol{\theta}$ is a $k \times 1$ parameter vector, and $\mathbf{N}$ is an $n \times 1$ random vector whose components N_i, $i = 1, \ldots, n$ are the errors or noise associated with the ith observation Y_i. We shall assume without loss of generality that $E[\mathbf{N}] = \mathbf{0}$.†

Equation 6.5–2 is known as the *linear model*. We now ask the following question: How do we extract a "good" estimate of $\boldsymbol{\theta}$ from the observed values of $\mathbf{Y}$ if we restrict our estimator $\hat{\boldsymbol{\Theta}}$ be a linear function of $\mathbf{Y}$? By a linear function we mean

$$\hat{\boldsymbol{\Theta}} = \mathbf{B}\mathbf{Y}, \tag{6.5–3}$$

where $\mathbf{B}$, which *does not* depend on $\mathbf{Y}$, is to be determined. The problem posed here is of great practical significance. It is one of the most fundamental problems in parameter estimation theory and covered in great detail in numerous books, for example, Kendall and Stuart [6–2] and Lewis and Odell [6–7]. It also is an immediate application of the probability theory of random vectors and is f for understanding numerous topics in the second half of this book Chapter 11.

Before computing the matrix $\mathbf{B}$ in Equation 6.5–3 for various cas first develop some results from matrix calculus.

Derivative of a scalar with respect to a vector. Let $q(\mathbf{x})$ be a sca of the vector $\mathbf{x} = (x_1, \ldots, x_n)^T$. Then

$$\frac{dq(\mathbf{x})}{d\mathbf{x}} \triangleq \left(\frac{\partial q}{\partial x_1}, \ldots, \frac{\partial q}{\partial x_n}\right)^T.$$

Thus the derivative of $q(\mathbf{x})$ with respect to $\mathbf{x}$ is a *column vector* whose nent is the partial derivative of $q(\mathbf{x})$ with respect to x_i.

Derivative of quadratic forms. Let $\mathbf{A}$ be a real-symmetric $n \times n$ ma $\mathbf{x}$ be an arbitrary n-vector. Then the derivative of the quadratic form

$$q(\mathbf{x}) \triangleq \mathbf{x}^T\mathbf{A}\mathbf{x}$$

with respect to $\mathbf{x}$ is

$$\frac{dq(\mathbf{x})}{d\mathbf{x}} = 2\mathbf{A}\mathbf{x}.$$

The proof of Equation 6.5-5 is obtained by writing

$$q(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j$$

$$= \sum_{i=1}^{n} x_i^2 a_{ii} + \sum_{i\neq j}^{n}\sum^{n} a_{ij} x_i x_j.$$

Hence

$$\frac{\partial q(\mathbf{x})}{\partial x_k} = 2x_k a_{kk} + 2\sum_{i\neq k} a_{ki} x_i$$

$$= 2\sum_{i=1}^{n} a_{ki} x_i$$

or

$$\frac{dq(\mathbf{x})}{d\mathbf{x}} = 2\mathbf{A}\mathbf{x}.$$

† The symbol $\mathbf{0}$ here stands for the zero vector, that is, the vector whose components are all zero.

Derivative of scalar products.† Let $\mathbf{a}$ and $\mathbf{x}$ be two n-vectors. Then with $y = \mathbf{a}^T\mathbf{x}$, we obtain

$$\frac{dy}{d\mathbf{x}} = \mathbf{a}. \tag{6.5–7}$$

Let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{A}$ be two n-vectors and an $n \times n$ matrix respectively. Then with $q \triangleq \mathbf{y}^T\mathbf{A}\mathbf{x}$,

$$\frac{\partial q}{\partial \mathbf{x}} = \mathbf{A}^T\mathbf{y}. \tag{6.5–8}$$

We return now to Equation 6.5–2:

$$\mathbf{Y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{N}$$

and assume that (recall $E[\mathbf{N}] = \mathbf{0}$)

$$\mathbf{K} \triangleq E[\mathbf{N}\mathbf{N}^T] = \sigma^2\mathbf{I} \tag{6.5–9}$$

where $\mathbf{I}$ is the identity matrix. Equation 6.5–9 is equivalent to stating that the measurement errors N_i, that is, $i = 1, \ldots, n$ are uncorrelated, and their variances are the same and equal to σ^2. This situation is sometimes called *white* noise.

We have not yet defined what we mean by a *good* or *best* estimator of θ. A reasonable choice is to find a $\hat{\boldsymbol{\Theta}}$ that *minimizes* the sum squares S defined by

$$S \triangleq (\mathbf{Y} - \mathbf{H}\hat{\boldsymbol{\Theta}})^T(\mathbf{Y} - \mathbf{H}\hat{\boldsymbol{\Theta}}) \triangleq \|\mathbf{Y} - \mathbf{H}\hat{\boldsymbol{\Theta}}\|^2. \tag{6.5–10}$$

Note that by finding $\hat{\boldsymbol{\Theta}}$ that best fits the measurement $\mathbf{Y}$ in the sense of minimizing $\|\mathbf{Y} - \mathbf{H}\hat{\boldsymbol{\Theta}}\|^2$, we are realizing what is commonly called a *least-squares* fit to the data. For this reason finding $\hat{\boldsymbol{\Theta}}$ that minimizes S in Equation 6.5–10 is called the least-squares (LS) method. To find the minimum of S with respect to $\hat{\boldsymbol{\Theta}}$, write

$$S = \mathbf{Y}^T\mathbf{Y} + \hat{\boldsymbol{\Theta}}^T\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}^T\mathbf{H}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{H}\hat{\boldsymbol{\Theta}}$$

and compute

$$\frac{\partial S}{\partial \hat{\boldsymbol{\Theta}}} = 0 = 2[\mathbf{H}^T\mathbf{H}]\hat{\boldsymbol{\Theta}} - 2\mathbf{H}^T\mathbf{Y},$$

† The *scalar or inner product* of two n-vectors, say $\mathbf{a}$ and $\mathbf{b}$, is $\mathbf{a}^T\mathbf{b}(= \mathbf{b}^T\mathbf{a})$. For the relation between scalar product and norm see Section 5.3. For the definitions of scalar product and norm of square-integrable functions see Section 4.4.

whence (assuming $\mathbf{H}^T\mathbf{H}$ has an inverse)

$$\hat{\boldsymbol{\Theta}}_{LS} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Y}.$$

Comparing our result with Equation 6.5–3 we see that $\mathbf{B}$ in E given by $\mathbf{B}_0 \triangleq (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ in the LS method. Equation 6.5–11 is of $\boldsymbol{\theta}$ based on the measurement $\mathbf{Y}$.

The astute reader will have noticed that we never involved $\sigma^2\mathbf{I}$. Indeed, in arriving at Equation 6.5–11 we essentially treated tic and merely obtained $\hat{\boldsymbol{\Theta}}_{LS}$ as the *generalized inverse* (see Lewi p. 6]) of the system of equations $\mathbf{Y} = \mathbf{H}\boldsymbol{\theta}$. As it stands, the estim Equation 6.5–11 has no claim to being optimum. However, whe of the noise $\mathbf{N}$ is as in Equation 6.5–9 then $\hat{\boldsymbol{\Theta}}_{LS}$ does indeed have ties in an important sense. However, before discussing this poi tion 6.6, we give some examples.

Example 6.5–1:

We are given the following data

$$6.2 = 3\theta + n_1$$
$$7.8 = 4\theta + n_2$$
$$2.2 = \theta + n_3.$$

Find the LS estimate of θ.

Solution: The data can be put in the form

$$\mathbf{y} = \mathbf{H}\theta + \mathbf{n},$$

where $\mathbf{y} = (6.2, 7.8, 2.2)^T$ is a realization of $\mathbf{Y}$, $\mathbf{H} = (3, 4, 1)^T$ and a realization of $\mathbf{N}$. Hence $\mathbf{H}^T\mathbf{H} = \sum H_i^2 = 26$ and $\mathbf{H}^T\mathbf{y} = \sum_{i=1}^{3} H$

$$\hat{\theta}_{LS} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} = \frac{\sum_{i=1}^{3} H_i y_i}{\sum_{i=1}^{3} H_i^2} = \frac{52}{26} = 2.$$

Example 6.4–2:

(Reference 6–2, p. 77.) Let $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ be a two-component pa be estimated, and let $\mathbf{H}$ be a $n \times 2$ matrix of coefficients partiti vectors as $\mathbf{H} = (\mathbf{H}_1\mathbf{H}_2)$ where $\mathbf{H}_i$, $i = 1, 2$ is an n-vector. Then w

representing the observation data, the linear model assumes the form

$$\mathbf{Y} = (\mathbf{H}_1\mathbf{H}_2)\boldsymbol{\theta} + \mathbf{N}$$

and the LS estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\Theta}}_{\text{LS}} = \begin{bmatrix} \mathbf{H}_1^T\mathbf{H}_1 & \mathbf{H}_1^T\mathbf{H}_2 \\ \mathbf{H}_2^T\mathbf{H}_1 & \mathbf{H}_2^T\mathbf{H}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{H}_1^T\mathbf{Y} \\ \mathbf{H}_2^T\mathbf{Y} \end{bmatrix}.$$

6.6 OPTIMAL PROPERTIES OF LEAST-SQUARES ESTIMATORS; THE GAUSS-MARKOV THEOREM

The LS estimator has a number of properties that account for its widespread use in estimation problems. It is simple to construct, does not require knowledge of the pdf of $\mathbf{N}$, is unbiased, and has a minimum variance property, which we discuss below.

Unbiasedness of the LS Estimator

To show that $\hat{\boldsymbol{\Theta}}_{\text{LS}}$ in Equation 6.5–11 is unbiased we must show that $E[\hat{\boldsymbol{\Theta}}_{\text{LS}}] = \boldsymbol{\theta}$. We write

$$\begin{aligned} \hat{\boldsymbol{\Theta}}_{\text{LS}} &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Y} \\ &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T(\mathbf{H}\boldsymbol{\theta} + \mathbf{N}) \\ &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} + (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{N} \\ &= \boldsymbol{\theta} + \mathbf{B}_0\mathbf{N} \quad (\mathbf{B}_0 \triangleq (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T). \end{aligned} \tag{6.6–1}$$

Hence

$$E[\hat{\boldsymbol{\Theta}}_{\text{LS}}] = \boldsymbol{\theta} + \mathbf{B}_0E[\mathbf{N}] = \boldsymbol{\theta}$$

since, by assumption, $E[\mathbf{N}] = \mathbf{0}$.

The covariance of $\hat{\boldsymbol{\Theta}}_{\text{LS}}$ is easily computed from Equation 6.6–1. Thus

$$\begin{aligned} E[(\hat{\boldsymbol{\Theta}}_{\text{LS}} - \boldsymbol{\theta})(\hat{\boldsymbol{\Theta}}_{\text{LS}} - \boldsymbol{\theta})^T] &= E[\mathbf{B}_0\mathbf{N}\mathbf{N}^T\mathbf{B}_0^T] \\ &= \mathbf{B}_0E[\mathbf{N}\mathbf{N}^T]\mathbf{B}_0^T \\ &= \sigma^2\mathbf{B}_0\mathbf{B}_0^T \quad (\text{since } E[\mathbf{N}\mathbf{N}^T] = \sigma^2\mathbf{I}) \\ &= \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}. \end{aligned} \tag{6.6–2}$$

Minimum Variance Property of $\hat{\boldsymbol{\Theta}}_{\text{LS}}$

We now demonstrate one of the most important properties of $\hat{\boldsymbol{\Theta}}_{\text{LS}}$, namely, its minimum variance property. Because of this property, the LS estimator is sometimes considered a best estimator and given the acronym BLUE (Best L Unbiased Estimator).

To begin with, consider any linear unbiased estimator $\hat{\boldsymbol{\Theta}} = (\hat{\Theta}_1, \ldots, \hat{\Theta}$ $\boldsymbol{\theta}$ of the form $\hat{\boldsymbol{\Theta}} = \mathbf{BY}$. If $\hat{\boldsymbol{\Theta}}$ is to be unbiased then

$$\begin{aligned} E[\hat{\boldsymbol{\Theta}}] &= E[\mathbf{BY}] = E[\mathbf{B}(\mathbf{H}\boldsymbol{\theta} + \mathbf{N})] \\ &= \mathbf{BH}\boldsymbol{\theta} \\ &= \boldsymbol{\theta} \quad (\text{by unbiasedness of } \hat{\boldsymbol{\Theta}}). \end{aligned}$$

The result $\mathbf{BH}\boldsymbol{\theta} = \boldsymbol{\theta}$ implies that

$$\mathbf{BH} = \mathbf{I}. \tag{6}$$

Equation 6.6–3 is a necessary and sufficient condition for $\mathbf{BY}$ to be an unb estimator of $\boldsymbol{\theta}$. Now consider the covariance matrix $\mathbf{K}_{\hat{\Theta}}$ of the error $\hat{\boldsymbol{\Theta}} -$

$$\begin{aligned} \mathbf{K}_{\hat{\Theta}} &= E[(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta})^T] \\ &= \sigma^2\mathbf{B}\mathbf{B}^T. \end{aligned} \tag{6}$$

To arrive at Equation 6.6–4 we used the facts that (1) $\mathbf{BH} = \mathbf{I}$, (2) that $\hat{\boldsymbol{\Theta}} = \boldsymbol{\theta}$ and (3) that $\mathbf{K} \triangleq E[\mathbf{N}\mathbf{N}^T] = \sigma^2\mathbf{I}$. We now ask what matrix $\mathbf{B}$ will minimi diagonal elements of Equation 6.6–4, which are the variance $\sigma^2_{\hat{\Theta}_i}$ of the estin $\hat{\Theta}_i$, $i = 1, \ldots, n$. We can solve this problem by using the following ident

$$\mathbf{B}\mathbf{B}^T = \mathbf{B}_0\mathbf{B}_0^T + (\mathbf{B} - \mathbf{B}_0)(\mathbf{B} - \mathbf{B}_0)^T. \tag{6}$$

To prove Equation 6.6–5 we need only substitute $\mathbf{B}_0 = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ and u fact that $\mathbf{BH} = \mathbf{I}$. Indeed, observing that

$$\mathbf{B}_0\mathbf{B}_0^T = (\mathbf{H}^T\mathbf{H})^{-1}$$

and

$$\mathbf{B}\mathbf{B}_0^T = (\mathbf{H}^T\mathbf{H})^{-1} = \mathbf{B}_0\mathbf{B}^T$$

and expanding Equation 6.6–5 to

$$\mathbf{B}\mathbf{B}^T = 2\mathbf{B}_0\mathbf{B}_0^T + \mathbf{B}\mathbf{B}^T - \mathbf{B}_0\mathbf{B}^T - \mathbf{B}\mathbf{B}_0^T \tag{6}$$

enables us to demonstrate the identity.

Finding the minimum variance estimator is readily accomplished o realize that for any real matrix $\mathbf{A}$ the diagonal terms of $\mathbf{A}\mathbf{A}^T$ are always s squares and hence nonnegative. In Equation 6.6–5 the diagonal terms of $\mathbf{B}$

minimized when the diagonal terms of $(\mathbf{B} - \mathbf{B}_0)(\mathbf{B} - \mathbf{B}_0)^T$ are minimized. But the smallest that the latter can get is zero since they are sums of squares. Thus $\mathbf{BB}^T$ has strictly minimum diagonal elements when $\text{tr}[(\mathbf{B} - \mathbf{B}_0)(\mathbf{B} - \mathbf{B}_0)^T] = 0$, which is achieved only when $\mathbf{B} = \mathbf{B}_0$. Hence we have shown that the LS estimator

$$\hat{\boldsymbol{\Theta}}_{\text{LS}} = \mathbf{B}_0\mathbf{Y} \quad (\mathbf{B}_0 \triangleq (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T) \tag{6.6–7}$$

is also the *minimum variance*, unbiased linear estimator of θ. Henceforth such estimators will have the subscript 0 rather than LS to emphasize their optimum, that is, minimum-variance property. The fact that the LS estimator is minimum variance when $\mathbf{K} = \sigma^2\mathbf{I}$ is a special case of the Gauss-Markov theorem. A slightly more general form of this special case is given below.

Theorem 6.6–1. Consider the model

$$\mathbf{Y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{N},$$

where $E[\mathbf{N}] = \mathbf{0}$ $E[\mathbf{N}\mathbf{N}^T] \triangleq \mathbf{K} = \sigma^2\mathbf{I}$. Suppose we want to estimate a linear function of $\boldsymbol{\theta}$, say $\boldsymbol{\phi} = \mathbf{D}\boldsymbol{\theta}$ where $\mathbf{D}$ is a matrix of known coefficients. Then the minimum-variance, unbiased, linear estimator $\hat{\boldsymbol{\Phi}}_0$ of $\boldsymbol{\phi}$ is given by

$$\hat{\boldsymbol{\Phi}}_0 = \mathbf{D}[\mathbf{H}^T\mathbf{H}]^{-1}\mathbf{H}^T\mathbf{Y}. \tag{6.6–8}$$

Comment. We have already proven the validity of Equation 6.6–8 for $\mathbf{D} = \mathbf{I}$ in the discussion leading up to Equation 6.6–7. The demonstration of Equation 6.6–8 for $\mathbf{D}$ arbitrary is a straightforward extension of the case when $\mathbf{D} = \mathbf{I}$. We leave this as well as the demonstration that any linear unbiased estimator of $\boldsymbol{\phi}$ of the form $\boldsymbol{\Phi} = \mathbf{LY}$ where $\mathbf{L}$ is to be determined must satisfy

$$\mathbf{LH} = \mathbf{D} \tag{6.6–9}$$

as exercises to the reader. ■

So far we have assumed that the noise $\mathbf{N}$ has covariance matrix $\mathbf{K} = \sigma^2\mathbf{I}$. What is the minimum variance, unbiased, linear estimator in the general case, that is, when $\mathbf{K}$ is an arbitrary covariance matrix? The answer follows.

Consider again the model

$$\mathbf{Y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{N}, \tag{6.6–10}$$

where all quantities are defined as before including $E[\mathbf{N}] = \mathbf{0}$, except the covariance matrix $\mathbf{K}$ of $\mathbf{N}$ is arbitrary positive definite. Then from Equation 5.6–14 we know that there exists a factorization (we change the notation slightly from Eq. 5.6–14):

$$\mathbf{K}^{-1} = \mathbf{CC}^T \tag{6.6–11}$$

such that $\mathbf{C}^T\mathbf{KC} = \mathbf{I}$ and we known how to compute $\mathbf{C}$. Now consider the transformation

$$\mathbf{W} = \mathbf{C}^T\mathbf{Y}. \tag{6.6–12}$$

Then premultiplying Equation 6.6–10 by $\mathbf{C}^T$ yields the matrix equation

$$\begin{aligned}\mathbf{W} &= \mathbf{C}^T\mathbf{H}\boldsymbol{\theta} + \mathbf{C}^T\mathbf{N} \\ &= \mathbf{H}'\boldsymbol{\theta} + \mathbf{N}'.\end{aligned} \tag{6.6–13}$$

where $\mathbf{H}' \triangleq \mathbf{C}^T\mathbf{H}$ and $\mathbf{N}' = \mathbf{C}^T\mathbf{N}$. Now observe that

$$\begin{aligned}\mathbf{K}_{\mathbf{N}'} &= E[\mathbf{N}'\mathbf{N}'^T] \\ &= E[\mathbf{C}^T\mathbf{N}\mathbf{N}^T\mathbf{C}] \\ &= \mathbf{C}^T E[\mathbf{N}\mathbf{N}^T]\mathbf{C} \\ &= \mathbf{C}^T\mathbf{K}\mathbf{C} \\ &= \mathbf{I}, \text{ the identity matrix.}\end{aligned} \tag{6.6–14}$$

Hence from Equation 6.6–7, the LS, minimum-variance, unbiased, linear estimator of θ is given by

$$\begin{aligned}\hat{\boldsymbol{\Theta}}_0 &= (\mathbf{H}'^T\mathbf{H}')^{-1}\mathbf{H}'^T\mathbf{W} \\ &= (\mathbf{H}^T\mathbf{C}\mathbf{C}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}\mathbf{C}^T\mathbf{Y} \\ &= (\mathbf{H}^T\mathbf{K}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{K}^{-1}\mathbf{Y}.\end{aligned} \tag{6.6–15}$$

Equation 6.6–15 is what is generally taken to be as the statement of the celebrated Gauss-Markov theorem. To show that $\hat{\boldsymbol{\Theta}}_0$ is unbiased we write

$$\begin{aligned}E[\hat{\boldsymbol{\Theta}}_0] &= E[(\mathbf{H}^T\mathbf{K}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{K}^{-1}(\mathbf{H}\boldsymbol{\theta} + \mathbf{N})] \\ &= \boldsymbol{\theta}.\end{aligned}$$

The covariance matrix of $\hat{\boldsymbol{\Theta}}_0$ is

$$\begin{aligned}E[(\hat{\boldsymbol{\Theta}}_0 - \boldsymbol{\theta})(\hat{\boldsymbol{\Theta}}_0 - \boldsymbol{\theta})^T] &= (\mathbf{H}^T\mathbf{K}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{K}^{-1}\mathbf{K} \\ &\quad \times \mathbf{K}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{K}^{-1}\mathbf{H})^{-1} \\ &= (\mathbf{H}^T\mathbf{K}^{-1}\mathbf{H})^{-1}\end{aligned} \tag{6.6–1}$$

and from the theory advanced earlier, its diagonal elements, that is, the variance $\sigma^2_{\hat{\Theta}_i}$ of the individual estimators $\hat{\Theta}_i$, $i = 1, \ldots, n$ in $\hat{\boldsymbol{\Theta}} \triangleq (\hat{\Theta}_1, \ldots, \hat{\Theta}_n)$ a minimal. Let us summarize this result in the following theorem.

Theorem 6.6–2. (The Gauss-Markov theorem.) Consider the model

$$\mathbf{Y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{N},$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations, $\mathbf{H}$ is an $n \times k$ $(n > k)$ matrix of known coefficients, $\boldsymbol{\theta}$ is a $k \times 1$ vector of parameters, and $\mathbf{N}$ is an $n \times 1$ random vector consisting of "measurement" noise with

$$E[\mathbf{N}] = \mathbf{0}$$

and

$$E[\mathbf{N}\mathbf{N}^T] \triangleq \mathbf{K}.$$

Then the minimum-variance, unbiased, linear estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\Theta}}_0 = (\mathbf{H}^T\mathbf{K}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{K}^{-1}\mathbf{Y}. \quad \blacksquare$$

Extension. The minimum-variance, unbiased, linear estimate of a linear function of $\boldsymbol{\theta}$, that is, $\boldsymbol{\phi} = \mathbf{D}\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\Phi}}_0 = \mathbf{D}(\mathbf{H}^T\mathbf{K}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{K}^{-1}\mathbf{Y}. \tag{6.6–17}$$

The proof of Equation 6.6–17 follows directly from the facts that (1) any unbiased estimator $\hat{\boldsymbol{\Phi}}$ of $\boldsymbol{\phi}$ that is a linear function of $\mathbf{Y}$, that is, $\hat{\boldsymbol{\Phi}} = \mathbf{L}\mathbf{Y}$ must satisfy

$$\mathbf{L}\mathbf{H} = \mathbf{D}$$

for unbiasedness and (2) the following identity with $\mathbf{L}_0 \triangleq \mathbf{D}(\mathbf{H}^T\mathbf{K}_N^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{K}_N^{-1}$ holds:

$$\mathbf{L}\mathbf{L}^T = \mathbf{L}_0\mathbf{L}_0^T + (\mathbf{L} - \mathbf{L}_0)(\mathbf{L} - \mathbf{L}_0)^T. \tag{6.6–18}$$

We leave the details as an exercise to the reader.

6.7 ESTIMATION OF RANDOM VARIABLES

In the previous sections we considered the estimation of a parameter θ (or $\boldsymbol{\theta}$) by observing random variables and forming a function of these random variables called an estimator. This estimator was then used to estimate the unknown parameter. We now consider a different problem, namely that of estimating one random variable with another or one random vector with another. We introduce the basic ideas in this section; subsequent development and application of these i be taken up in Chapter 11.

To make clear what we mean by estimating one random variable other, consider the following example: Let X_1 denote the barometric pres and X_2 denote the rate of change of the BP at $t = 0$. Let Y denote th humidity one hour after measuring $\mathbf{X} = (X_1, X_2)^T$. Clearly, in this case, are dependent r.v.'s; then using $\mathbf{X}$ to estimate Y is a case of estimating with another (actually $\mathbf{X}$ here is a random vector).†

In terms of the axiomatic theory we can describe the problem of e one r.v. with another in the following terms: Consider an underlying expe with probability space $\mathcal{P} \triangleq (\Omega, \mathcal{F}, P)$. Let X and Y be two r.v.'s defined every $\zeta \in \Omega$, we generate the numbers $X(\zeta)$, $Y(\zeta)$. Suppose we can obs $X(\zeta)$; how do we proceed to estimate $Y(\zeta)$ in some optimum fashion?

At this point the reader may wonder why observing $X(\zeta)$ doesn't specify $Y(\zeta)$. After all, since X and Y are deterministic functions, why reason that $X(\zeta)$ specifies ζ specifies $Y(\zeta)$? The answer is that observing not, in general, uniquely specify the outcome ζ and therefore does not specify $Y(\zeta)$. For example, let $\Omega = \{-2, -1, 0, 1, 2\}$, $X(\zeta) \triangleq \zeta^2$ and $Y(\zeta)$ the observation $X(\zeta) = 4$ is associated with the outcomes $\zeta = 2$ *or* ζ course, these may not be equally probable) and $Y(2) = 2$ while $Y(-$ Hence all we can say about $Y(\zeta)$ after observing $X(\zeta)$ is that $Y(\zeta)$ has -2. If all outcomes $\zeta \in \Omega$ are equally likely then the *a priori* probability is $\frac{1}{5}$ and $P[Y = 2 \mid X = 4] = \frac{1}{2}$.

Assume, at first, for simplicity that we are constrained to estimate linear function aX. Assume $E[X] = E[Y] = 0$. Note that a generalizati problem has already been treated in Example 4.3–4. The mean squ (MSE) in estimating Y by aX is given by

$$\begin{aligned}\varepsilon &\triangleq E[(Y - aX)^2] \\ &= \sigma_Y^2 - 2a\,\mathrm{Cov}(X, Y) + a^2\sigma_X^2\end{aligned}$$

Setting the first derivative with respect to a equal to zero to find the min obtain

$$a_o = \frac{\mathrm{Cov}(X, Y)}{\sigma_X^2}.$$

Equation 6.7–2 furnishes the value of a, which yields a minimum me error (MMSE) if we restrict ourselves to linear estimates of the form Y

† The abbreviation r.v. can mean random variable or random vector without amb

‡ All random variables in this section are initially assumed to be real. However discussion we shall generalize to include complex r.v.'s.