2

Probability Concepts in Least Squares

The excitement that a gambler feels when making a bet is equal to the amount he might win times the probability of winning it. Pascal, Blaise

THE intuitively reasonable *principle of least squares* was put forth in §1.2 and employed as the starting point for all developments of Chapter 1. In the present chapter, several alternative paths are followed to essentially the same mathematical conclusions as Chapter 1. The primary function of the present chapter is to place the results of Chapter 1 upon a more rigorous (or at least, a better understood) foundation. A number of new and computationally most useful extensions of the estimation results of Chapter 1 come from the developments shown herein. In particular, minimal variance estimation and maximum likelihood estimation will be explored, and a connection to the least squares problem will be shown. Using these estimation techniques, the elusive weight matrix will be rigorously identified as the inverse of the measurement-error covariance matrix, and some most important *nonuniqueness properties* developed in §2.5. Methods for rigorously accounting for *a priori* parameter estimates and their uncertainty will also be developed. Finally, many other useful concepts will be explored, including: unbiased estimates and the Cramér-Rao inequality; other advanced topics such as Bayesian estimation, analysis of covariance errors, and ridge estimation are introduced as well. These concepts are useful for the analysis of least squares estimation by incorporating probabilistic approaches.

Familiarity with basic concepts in probability is necessary for comprehension of the material in the present chapter. Should the reader anticipate or encounter difficulty in the following developments, Appendix B provides an adequate review of the concepts needed herein.

2.1 Minimum Variance Estimation

Here we introduce one of the most important and useful concepts in estimation. Minimum variance estimation can give the "best way" (in a probabilistic sense) to find the optimal estimates. First, a minimum variance estimator is derived without *a priori* estimates. Then these results are extended to the case where *a priori* estimates are given.

2.1.1 Estimation without *a priori* State Estimates

As in Chapter 1, we assume a linear observation model

$$\overset{(m\times 1)}{\tilde{\mathbf{y}}} = \overset{(m\times n)}{H}\overset{(n\times 1)}{\mathbf{x}} + \overset{(m\times 1)}{\mathbf{v}} \tag{2.1}$$

We desire to estimate $\mathbf{x}$ as a linear combination of the measurements $\tilde{\mathbf{y}}$ as

$$\overset{(n\times 1)}{\hat{\mathbf{x}}} = \overset{(n\times m)}{M}\overset{(m\times 1)}{\tilde{\mathbf{y}}} + \overset{(n\times 1)}{\mathbf{n}} \tag{2.2}$$

An "optimum" choice of the quantities M and $\mathbf{n}$ is sought. The minimum variance definition of "optimum" M and $\mathbf{n}$ is that the variance of *all* n estimates, $\hat{x}_i$, from their respective "true" values is minimized:*

$$J_i = \frac{1}{2}E\left\{\left(\hat{x}_i - x_i\right)^2\right\}, \quad i = 1, 2, \ldots, n \tag{2.3}$$

This clearly requires n minimizations depending upon the same M and $\mathbf{n}$; it may not be clear at this point that the problem is well-defined and whether or not M and $\mathbf{n}$ exist (or can be found if they do exist) to accomplish these n minimizations.

If the linear model (2.1) is strictly valid, then, for the special case of perfect measurements $\mathbf{v} = 0$ the model (2.1) should be exactly satisfied by the perfect measurements $\mathbf{y}$ and the true state $\mathbf{x}$ as

$$\tilde{\mathbf{y}} \equiv \mathbf{y} = H\mathbf{x} \tag{2.4}$$

An obvious requirement upon the desired estimator (2.2) is that perfect measurements should result (if a solution is possible) when $\hat{\mathbf{x}} = \mathbf{x} =$ true state. Thus, this requirement can be written by substituting $\hat{\mathbf{x}} = \mathbf{x}$ and $\tilde{\mathbf{y}} = H\mathbf{x}$ into eqn. (2.2) as

$$\mathbf{x} = MH\mathbf{x} + \mathbf{n} \tag{2.5}$$

We conclude that M and $\mathbf{n}$ satisfy the constraints

$$\mathbf{n} = \mathbf{0} \tag{2.6}$$

and

$$MH = I \tag{2.7a}$$

$$H^T M^T = I \tag{2.7b}$$

Equation (2.6) is certainly useful information! The desired estimator then has the form

$$\hat{\mathbf{x}} = M\tilde{\mathbf{y}} \tag{2.8}$$

We are now concerned with determining the optimum choice of M which accomplishes the n minimizations of (2.3), subject to the constraint (2.7).

* $E\{\ \}$ denotes "*expected value*" of $\{\ \}$, see Appendix B.

Subsequent manipulations will be greatly facilitated by partitioning the various matrices as follows: The unknown M-matrix is partitioned by rows as

$$M = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_n \end{bmatrix}, \quad M_i \equiv \{M_{i1}\; M_{i2} \cdots M_{im}\} \tag{2.9}$$

or

$$M^T = \begin{bmatrix} M_1^T \; M_2^T \; \cdots \; M_n^T \end{bmatrix} \tag{2.10}$$

The identity matrix can be partitioned by rows and columns as

$$I = \begin{bmatrix} I_1^r \\ I_2^r \\ \vdots \\ I_n^r \end{bmatrix} = \begin{bmatrix} I_1^c \; I_2^c \; \cdots \; I_n^c \end{bmatrix}, \quad \text{note } I_i^r = (I_i^c)^T \tag{2.11}$$

The constraint in eqn. (2.7) can now be written as

$$H^T M_i^T = I_i^c, \quad i = 1, 2, \dots, n \tag{2.12a}$$

$$M_i H = I_i^r, \quad i = 1, 2, \dots, n \tag{2.12b}$$

and the i^{th} element of $\hat{\mathbf{x}}$ from eqn. (2.8) can be written as

$$\hat{x}_i = M_i \tilde{\mathbf{y}}, \quad i = 1, 2, \dots, n \tag{2.13}$$

A glance at eqn. (2.13) reveals that $\hat{x}_i$ depends *only* upon the elements of M contained in the i^{th} row. A similar statement holds for the constraint equations (2.12); the elements of the i^{th} row are independently constrained. This "uncoupled" nature of eqns. (2.12) and (2.13) is the key feature which allows one to carry out the n "separate" minimizations of eqn. (2.3).

The i^{th} variance (2.3) to be minimized, upon substituting eqn. (2.13), can be written as

$$J_i = \frac{1}{2} E\left\{ \left(M_i \tilde{\mathbf{y}} - x_i\right)^2 \right\}, \quad i = 1, 2, \dots, n \tag{2.14}$$

Substituting the observation from eqn. (2.1) into eqn. (2.14) yields

$$J_i = \frac{1}{2} E\left\{ (M_i H \mathbf{x} + M_i \mathbf{v} - x_i)^2 \right\}, \quad i = 1, 2, \dots, n \tag{2.15}$$

Incorporating the constraint equations from eqn. (2.12) into eqn. (2.15) yields

$$J_i = \frac{1}{2} E\left\{ \left(I_i^r \mathbf{x} + M_i \mathbf{v} - x_i\right)^2 \right\}, \quad i = 1, 2, \dots, n \tag{2.16}$$

But $I_i^r \mathbf{x} = x_i$, so that eqn. (2.16) reduces to

$$J_i = \frac{1}{2} E\left\{ (M_i \mathbf{v})^2 \right\}, \quad i = 1, 2, \dots, n \tag{2.17}$$

which can be rewritten as

$$J_i = \frac{1}{2} E\left\{ M_i \left(\mathbf{v}\,\mathbf{v}^T \right) M_i^T \right\}, \quad i = 1, 2, \ldots, n \tag{2.18}$$

But the only random variable on the right-hand side of eqn. (2.18) is $\mathbf{v}$; introducing the covariance matrix of measurement errors (assuming that $\mathbf{v}$ has zero mean, i.e., $E\{\mathbf{v}\} = 0$),

$$\text{cov}\{\mathbf{v}\} \equiv R = E\left\{ \mathbf{v}\,\mathbf{v}^T \right\} \tag{2.19}$$

then eqn. (2.18) reduces to

$$J_i = \frac{1}{2} M_i R M_i^T, \quad i = 1, 2, \ldots, n \tag{2.20}$$

The i^{th} constrained minimization problem can now be stated as: Minimize each of equations (2.20) subject to the corresponding constraint in eqn. (2.12). Using the method of Lagrange multipliers (Appendix C), the i^{th} augmented function is introduced as

$$J_i = \frac{1}{2} M_i R M_i^T + \boldsymbol{\lambda}_i^T \left(I_i^c - H^T M_i^T \right), \quad i = 1, 2, \ldots, n \tag{2.21}$$

where

$$\boldsymbol{\lambda}_i^T = \left\{ \lambda_{1_i}, \lambda_{2_i}, \ldots, \lambda_{n_i} \right\} \tag{2.22}$$

are n vectors of Lagrange multipliers.

The necessary conditions for eqn. (2.21) to be minimized are then

$$\nabla_{M_i^T} J_i = R M_i^T - H \boldsymbol{\lambda}_i = \mathbf{0}, \quad i = 1, 2, \ldots, n \tag{2.23}$$

$$\nabla_{\boldsymbol{\lambda}_i} J_i = I_i^c - H^T M_i^T = \mathbf{0}, \text{ or } M_i H = I_i^r, \quad i = 1, 2, \ldots, n \tag{2.24}$$

From eqn. (2.23), we obtain

$$M_i = \boldsymbol{\lambda}_i^T H^T R^{-1}, \quad i = 1, 2, \ldots, n \tag{2.25}$$

Substituting eqn. (2.25) into the second equation of eqn. (2.24) yields

$$\boldsymbol{\lambda}_i^T = I_i^r \left(H^T R^{-1} H \right)^{-1} \tag{2.26}$$

Therefore, substituting eqn. (2.26) into eqn. (2.25), the n rows of M are given by

$$M_i = I_i^r \left(H^T R^{-1} H \right)^{-1} H^T R^{-1}, \quad i = 1, 2, \ldots, n \tag{2.27}$$

It then follows that

$$M = \left(H^T R^{-1} H \right)^{-1} H^T R^{-1} \tag{2.28}$$

and the desired estimator (2.8) then has the final form

$$\boxed{\hat{\mathbf{x}} = \left(H^T R^{-1} H\right)^{-1} H^T R^{-1} \tilde{\mathbf{y}}} \tag{2.29}$$

which is referred to as the *Gauss-Markov Theorem*.

The minimal variance estimator (2.29) is identical to the least squares estimator (1.30), *provided that the weight matrix is identified as the inverse of the observation error covariance*. Also, the "sequential least squares estimation" results of §1.3 are seen to embody a special case "sequential minimal variance estimation;" it is simply necessary to employ R^{-1} as W in the sequential least squares formulation, but we still require R^{-1} to have the block diagonal structure assumed for W.

The previous derivation can also be shown in compact form, but requires using vector matrix differentiation. This is shown for completeness. We will see in §2.2 that the condition $MH = I$ gives an *unbiased* estimate of $\mathbf{x}$. Let us first define the error covariance matrix for an unbiased estimator, given by (see Appendix B for details)

$$P = E\left\{(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^T\right\} \tag{2.30}$$

We wish to determine M that minimizes eqn. (2.30) in some way. We will choose to minimize the trace of P since this is a common choice and intuitively makes sense. Therefore, applying this choice with the constraint $MH = I$ gives the following loss function to be minimized:

$$J = \frac{1}{2}\text{Tr}\left[E\left\{(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^T\right\}\right] + \text{Tr}[\Lambda(I - MH)] \tag{2.31}$$

where Tr denotes the trace operator, and Λ is an $n \times n$ matrix of Lagrange multipliers. We can also make use of the parallel axis theorem[1†] for an unbiased estimate (i.e., $MH = I$), which states that

$$E\left\{(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^T\right\} = E\left\{\hat{\mathbf{x}}\hat{\mathbf{x}}^T\right\} - E\{\mathbf{x}\}E\{\mathbf{x}\}^T \tag{2.32}$$

Substituting eqn. (2.1) into eqn. (2.8) leads to

$$\begin{aligned}\hat{\mathbf{x}} &= M\tilde{\mathbf{y}} \\ &= MH\mathbf{x} + M\mathbf{v}\end{aligned} \tag{2.33}$$

Next, taking the expectation of both sides of eqn. (2.33) and using $E\{\mathbf{v}\} = 0$ gives (note, $\mathbf{x}$ on the right-hand side of eqn. (2.33) is treated as a deterministic quantity)

$$E\left\{\hat{\mathbf{x}}\right\} = MH\mathbf{x} \tag{2.34}$$

[†] This terminology is actually more commonly used in analytical dynamics to determine the moment of inertia about some arbitrary axis, related by a parallel axis through the center of mass.[2, 3] However, in statistics the form of the equation is identical when taking second moments about an arbitrary random variable.

In a similar fashion, using $E\{\mathbf{v}\mathbf{v}^T\} = R$ and also assuming that $E\{\mathbf{x}\mathbf{v}^T\} = 0$ and $E\{\mathbf{v}\mathbf{x}^T\} = 0$ (i.e., $\mathbf{x}$ and $\mathbf{v}$ are assumed to be *uncorrelated*), we obtain

$$E\left\{\hat{\mathbf{x}}\hat{\mathbf{x}}^T\right\} = MH\mathbf{x}\mathbf{x}^T H^T M^T + MRM^T \tag{2.35}$$

Therefore, the loss function in eqn. (2.31) becomes

$$J = \frac{1}{2}\text{Tr}(MRM^T) + \text{Tr}[\Lambda(I - MH)] \tag{2.36}$$

Next, we will make use of the following useful trace identities (see Appendix A):

$$\frac{\partial}{\partial A}\text{Tr}(BAC) = B^T C^T \tag{2.37a}$$

$$\frac{\partial}{\partial A}\text{Tr}(ABA^T) = A(B + B^T) \tag{2.37b}$$

Thus, we have the following necessary conditions:

$$\nabla_M J = MR - \Lambda^T H^T = 0 \tag{2.38}$$

$$\nabla_\Lambda J = I - MH = 0 \tag{2.39}$$

Solving eqn. (2.38) for M yields

$$M = \Lambda^T H^T R^{-1} \tag{2.40}$$

Substituting eqn. (2.40) into eqn. (2.39), and solving for Λ^T gives

$$\Lambda^T = (H^T R^{-1} H)^{-1} \tag{2.41}$$

Finally, substituting eqn. (2.41) into eqn. (2.40) yields

$$M = (H^T R^{-1} H)^{-1} H^T R^{-1} \tag{2.42}$$

This is identical to the solution given by eqn. (2.28).

2.1.2 Estimation with *a priori* State Estimates

The preceding results will now be extended to allow rigorous incorporation of *a priori* estimates, $\hat{\mathbf{x}}_a$, of the state and associated *a priori* error covariance matrix Q. We again assume the linear observation model

$$\tilde{\mathbf{y}} = H\mathbf{x} + \mathbf{v} \tag{2.43}$$

and associated (assumed known) measurement-error covariance matrix

$$R = E\left\{\mathbf{v}\mathbf{v}^T\right\} \tag{2.44}$$

Suppose that the variable **x** is also unknown (i.e., it is now treated as a *random variable*). The *a priori* state estimates are given as the sum of the true state **x** and the errors in the *a priori* estimates **w**, so that

$$\hat{\mathbf{x}}_a = \mathbf{x} + \mathbf{w} \tag{2.45}$$

with associated (assumed known) *a priori* error covariance matrix

$$\text{cov}\{\mathbf{w}\} \equiv Q = E\left\{\mathbf{w}\,\mathbf{w}^T\right\} \tag{2.46}$$

where we assume that **w** has zero mean. We also assume that the measurement errors and *a priori* errors are uncorrelated so that $E\left\{\mathbf{w}\,\mathbf{v}^T\right\} = 0$.

We desire to estimate **x** as a linear combination of the measurements $\tilde{\mathbf{y}}$ and *a priori* state estimates $\hat{\mathbf{x}}_a$ as

$$\hat{\mathbf{x}} = M\tilde{\mathbf{y}} + N\hat{\mathbf{x}}_a + \mathbf{n} \tag{2.47}$$

An "optimum" choice of the M $(n \times m)$, N $(n \times n)$, and $\mathbf{n}$ $(n \times 1)$ matrices is desired. As before, we adopt the minimal variance definition of "optimum" to determine M, N, and $\mathbf{n}$ for which the variances of all n estimates, $\hat{x}_i$, from their respective true values, x_i, are minimized:

$$J_i = \frac{1}{2}E\left\{\left(\hat{x}_i - x_i\right)^2\right\}, \quad i = 1, 2, \ldots, n \tag{2.48}$$

If the linear model (2.43) is strictly valid, then for the special case of perfect measurements ($\mathbf{v} = \mathbf{0}$), the measurements **y** and the true state **x** should satisfy eqn. (2.43) exactly as

$$\mathbf{y} = H\mathbf{x} \tag{2.49}$$

If, in addition, the *a priori* state estimates are also perfect ($\hat{\mathbf{x}}_a = \mathbf{x}$, $\mathbf{w} = \mathbf{0}$), an obvious requirement upon the estimator in eqn. (2.47) is that it yields the true state as

$$\mathbf{x} = MH\mathbf{x} + N\mathbf{x} + \mathbf{n} \tag{2.50}$$

or

$$\mathbf{x} = (MH + N)\mathbf{x} + \mathbf{n} \tag{2.51}$$

Equation (2.51) indicates that M, N, and **n** must satisfy the constraints

$$\mathbf{n} = \mathbf{0} \tag{2.52}$$

and

$$MH + N = I \text{ or } H^T M^T + N^T = I \tag{2.53}$$

Because of eqn. (2.52), the desired estimator (2.47) has the form

$$\hat{\mathbf{x}} = M\tilde{\mathbf{y}} + N\hat{\mathbf{x}}_a \tag{2.54}$$

It is useful in subsequent developments to partition M, N, and I as

$$M = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_n \end{bmatrix}, \quad M^T = \begin{bmatrix} M_1^T \; M_2^T \; \cdots \; M_n^T \end{bmatrix} \tag{2.55}$$

$$N = \begin{bmatrix} N_1 \\ N_2 \\ \vdots \\ N_n \end{bmatrix}, \quad N^T = \begin{bmatrix} N_1^T \; N_2^T \; \cdots \; N_n^T \end{bmatrix} \tag{2.56}$$

and

$$I = \begin{bmatrix} I_1^r \\ I_2^r \\ \vdots \\ I_n^r \end{bmatrix} = \begin{bmatrix} I_1^c \; I_2^c \; \cdots \; I_n^c \end{bmatrix}, \quad I_i^r = (I_i^c)^T \tag{2.57}$$

Using eqns. (2.55), (2.56), and (2.57), the constraint equation (2.53) can be written as n independent constraints as

$$H^T M_i^T + N_i^T = I_i^c, \quad i = 1, 2, \ldots, n \tag{2.58a}$$

$$M_i H + N_i = I_i^r, \quad i = 1, 2, \ldots, n \tag{2.58b}$$

The i^{th} element of $\hat{\mathbf{x}}$, from eqn. (2.54), is

$$\hat{x}_i = M_i \tilde{\mathbf{y}} + N_i \hat{\mathbf{x}}_a, \quad i = 1, 2, \ldots, n \tag{2.59}$$

Note that both eqns. (2.58) and (2.59) depend *only* upon the elements of the i^{th} row, M_i, of M and the i^{th} row, N_i, of N. Thus the i^{th} variance (2.48) to be minimized is a function of the same $n+m$ unknowns (the elements of M_i and N_i) as is the i^{th} constraint, eqn. (2.58a) or eqn. (2.58b).

Substituting eqn. (2.59) into eqn. (2.48) yields

$$J_i = \frac{1}{2} E\left\{ \left(M_i \tilde{\mathbf{y}} + N_i \hat{\mathbf{x}}_a - x_i \right)^2 \right\}, \quad i = 1, 2, \ldots, n \tag{2.60}$$

Substituting eqns. (2.43) and (2.45) into eqn. (2.60) yields

$$J_i = \frac{1}{2} E\left\{ [(M_i H + N_i)\mathbf{x} + M_i \mathbf{v} + N_i \mathbf{w} - x_i]^2 \right\}, \quad i = 1, 2, \ldots, n \tag{2.61}$$

Making use of eqn. (2.58a), eqn. (2.61) becomes

$$J_i = \frac{1}{2} E\left\{ \left(I_i^r \mathbf{x} + M_i \mathbf{v} + N_i \mathbf{w} - x_i \right)^2 \right\}, \quad i = 1, 2, \ldots, n \tag{2.62}$$

Since $I_i^r \mathbf{x} = x_i$, eqn. (2.62) reduces to

$$J_i = \frac{1}{2} E\left\{(M_i \mathbf{v} + N_i \mathbf{w})^2\right\}, \quad i = 1, 2, \ldots, n \tag{2.63}$$

or

$$J_i = \frac{1}{2} E\left\{(M_i \mathbf{v})^2 + 2\,(M_i \mathbf{v})\,(N_i \mathbf{w}) + (N_i \mathbf{w})^2\right\}, \quad i = 1, 2, \ldots, n \tag{2.64}$$

which can be written as

$$\begin{aligned} J_i = \frac{1}{2} E\Big\{ & M_i \left(\mathbf{v}\,\mathbf{v}^T\right) M_i^T + 2 M_i \left(\mathbf{v}\,\mathbf{w}^T\right) N_i^T \\ & + N_i \left(\mathbf{w}\,\mathbf{w}^T\right) N_i^T \Big\}, \quad i = 1, 2, \ldots, n \end{aligned} \tag{2.65}$$

Therefore, using the defined covariances in eqns. (2.44) and (2.46), and since we have assumed that $E\left\{\mathbf{v}\,\mathbf{w}^T\right\} = 0$ (i.e., the errors are uncorrelated), eqn. (2.65) becomes

$$J_i = \frac{1}{2}\left[M_i R M_i^T + N_i Q N_i^T\right], \quad i = 1, 2, \ldots, n \tag{2.66}$$

The i^{th} minimization problem can then be restated as: Determine the M_i and N_i to minimize the i^{th} equation (2.66) subject to the constraint equation (2.53).

Using the method of Lagrange multipliers (Appendix C), the augmented functions are defined as

$$\begin{aligned} J_i = {} & \frac{1}{2}\left[M_i R M_i^T + N_i Q N_i^T\right] \\ & + \boldsymbol{\lambda}_i^T \left(I_i^c - H^T M_i^T - N_i^T\right), \quad i = 1, 2, \ldots, n \end{aligned} \tag{2.67}$$

where

$$\boldsymbol{\lambda}_i^T = \left\{\lambda_{1_i}, \lambda_{2_i}, \ldots, \lambda_{n_i}\right\} \tag{2.68}$$

is the i^{th} matrix of n Lagrange multipliers.

The necessary conditions for a minimum of eqn. (2.67) are

$$\nabla_{M_i^T} J_i = R M_i^T - H \boldsymbol{\lambda}_i = \mathbf{0}, \quad i = 1, 2, \ldots, n \tag{2.69}$$

$$\nabla_{N_i^T} J_i = Q N_i^T - \boldsymbol{\lambda}_i = \mathbf{0}, \quad i = 1, 2, \ldots, n \tag{2.70}$$

and

$$\nabla_{\boldsymbol{\lambda}_i} J_i = I_i^c - H^T M_i^T - N_i^T = \mathbf{0}, \quad i = 1, 2, \ldots, n \tag{2.71}$$

From eqns. (2.69) and (2.70), we obtain

$$M_i = \boldsymbol{\lambda}_i^T H^T R^{-1},\; M_i^T = R^{-1} H \boldsymbol{\lambda}_i, \quad i = 1, 2, \ldots, n \tag{2.72}$$

and

$$N_i = \boldsymbol{\lambda}_i^T Q^{-1},\; N_i^T = Q^{-1} \boldsymbol{\lambda}_i, \quad i = 1, 2, \ldots, n \tag{2.73}$$

Substituting eqns. (2.72) and (2.73) into (2.71) allows immediate solution for $\boldsymbol{\lambda}_i^T$ as

$$\boldsymbol{\lambda}_i^T = I_i^r \left(H^T R^{-1} H + Q^{-1}\right)^{-1}, \quad i = 1, 2, \ldots, n \tag{2.74}$$

Then substituting eqn. (2.74) into eqns. (2.72) and (2.73), the rows of M and N are

$$M_i = I_i^r \left(H^T R^{-1} H + Q^{-1}\right)^{-1} H^T R^{-1}, \quad i = 1, 2, \ldots, n \tag{2.75}$$

$$N_i = I_i^r \left(H^T R^{-1} H + Q^{-1}\right)^{-1} Q^{-1}, \quad i = 1, 2, \ldots, n \tag{2.76}$$

Therefore, the M and N matrices are

$$M = \left(H^T R^{-1} H + Q^{-1}\right)^{-1} H^T R^{-1} \tag{2.77}$$

$$N = \left(H^T R^{-1} H + Q^{-1}\right)^{-1} Q^{-1} \tag{2.78}$$

Finally, substituting eqns. (2.77) and (2.78) into eqn. (2.54) yields the minimum variance estimator

$$\boxed{\hat{\mathbf{x}} = \left(H^T R^{-1} H + Q^{-1}\right)^{-1} \left(H^T R^{-1} \tilde{\mathbf{y}} + Q^{-1} \hat{\mathbf{x}}_a\right)} \tag{2.79}$$

which allows rigorous processing of *a priori* state estimates $\hat{\mathbf{x}}_a$ and associated covariance matrices Q.

Notice the following limiting cases:

1. *A priori* knowledge very poor

$$\left(R \text{ finite}, \; Q \to \infty, \; Q^{-1} \to 0\right)$$

then eqn. (2.79) reduces immediately to the standard minimal variance estimator (2.29).

2. Measurements very poor

$$\left(Q \text{ finite}, \; R^{-1} \to 0\right)$$

then eqn. (2.79) yields $\hat{\mathbf{x}} = \hat{\mathbf{x}}_a$, an intuitively pleasing result!

Notice also that eqn. (2.79) can be obtained from the sequential least squares formulation of §1.3 by processing the *a priori* state information as a subset of the "observation" as follows: In eqns. (1.53) and (1.54) of the sequential estimation developments:

1. Set $\tilde{\mathbf{y}}_2 = \hat{\mathbf{x}}_a$, $H_2 = I$ (note: the dimension of $\tilde{\mathbf{y}}_2$ is n in this case), and $W_1 = R^{-1}$ and $W_2 = Q^{-1}$.

2. Ignore the "1" and "2" subscripts.

Then one immediately obtains eqn. (2.79).

We thus conclude that the minimal variance estimate (2.79) is in all respects consistent with the sequential estimation results of §1.3; to start the sequential process, one would probably employ the *a priori* estimates as

$$\begin{aligned}\hat{\mathbf{x}}_1 &= \hat{\mathbf{x}}_a \\ P_1 &= Q\end{aligned}$$

and process subsequent measurement subsets $\left\{\tilde{\mathbf{y}}_k,\, H_k,\, W_k\right\}$ with $W_k = R^{-1}$ for the minimal variance estimates of $\mathbf{x}$.

As in the case of estimation without *a priori* estimates, the previous derivation can also be shown in compact form. The following loss function to be minimized is

$$J = \frac{1}{2}\text{Tr}\left[E\left\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T\right\}\right] + \text{Tr}[\Lambda(I - MH - N)] \tag{2.80}$$

Substituting eqns. (2.43) and (2.45) into eqn. (2.54) leads to

$$\begin{aligned}\hat{\mathbf{x}} &= M\tilde{\mathbf{y}} + N\hat{\mathbf{x}}_a \\ &= (MH+N)\mathbf{x} + M\mathbf{v} + N\mathbf{w}\end{aligned} \tag{2.81}$$

Next, as before we assume that the true state $\mathbf{x}$ and error terms $\mathbf{v}$ and $\mathbf{w}$ are uncorrelated with each other. Using eqns. (2.44) and (2.46) with the uncorrelated assumption leads to

$$J = \frac{1}{2}\text{Tr}(MRM^T + NQN^T) + \text{Tr}[\Lambda(I - MH - N)] \tag{2.82}$$

Therefore, we have the following necessary conditions:

$$\nabla_M J = MR - \Lambda^T H^T = 0 \tag{2.83}$$

$$\nabla_N J = NQ - \Lambda^T = 0 \tag{2.84}$$

$$\nabla_\Lambda J = I - MH - N = 0 \tag{2.85}$$

Solving eqn. (2.83) for M yields

$$M = \Lambda^T H^T R^{-1} \tag{2.86}$$

Solving eqn. (2.84) for N yields

$$N = \Lambda^T Q^{-1} \tag{2.87}$$

Substituting eqns. (2.86) and (2.87) into eqn. (2.85), and solving for Λ^T gives

$$\Lambda^T = (H^T R^{-1} H + Q^{-1})^{-1} \tag{2.88}$$

Finally, substituting eqn. (2.88) into eqns. (2.86) and (2.87) yields

$$M = \left(H^T R^{-1} H + Q^{-1}\right)^{-1} H^T R^{-1} \tag{2.89}$$

$$N = \left(H^T R^{-1} H + Q^{-1}\right)^{-1} Q^{-1} \tag{2.90}$$

This is identical to the solutions given by eqns. (2.77) and (2.78).

2.2 Unbiased Estimates

The structure of eqn. (2.8) can also be used to prove that the minimal variance estimator is "unbiased." An estimator $\hat{\mathbf{x}}(\tilde{\mathbf{y}})$ is said to be an "unbiased estimator" of $\mathbf{x}$ if $E\left\{\hat{\mathbf{x}}(\tilde{\mathbf{y}})\right\} = \mathbf{x}$ for every possible value of $\mathbf{x}$.[4‡] If $\hat{\mathbf{x}}$ is biased, the difference $E\left\{\hat{\mathbf{x}}(\tilde{\mathbf{y}})\right\} - \mathbf{x}$ is called the "bias" of $\hat{\mathbf{x}}$. For the minimum variance estimate $\hat{\mathbf{x}}$, given by eqn. (2.29), to be unbiased M must satisfy the following condition:

$$MH = I \tag{2.91}$$

The proof of the unbiased condition is given by first substituting eqn. (2.1) into eqn. (2.13), leading to

$$\begin{aligned} \hat{\mathbf{x}} &= M\tilde{\mathbf{y}} \\ &= MH\mathbf{x} + M\mathbf{v} \end{aligned} \tag{2.92}$$

Next, taking the expectation of both sides of (2.92) and using $E\{\mathbf{v}\} = 0$ gives (again $\mathbf{x}$ on the right-hand side of eqn. (2.92) is treated as a deterministic quantity)

$$E\left\{\hat{\mathbf{x}}\right\} = MH\mathbf{x} \tag{2.93}$$

which gives the condition in eqn. (2.91). Substituting eqn. (2.28) into eqn. (2.91) shows that the estimator clearly produces an unbiased estimate of $\hat{\mathbf{x}}$.

The sequential least squares estimator can also be shown to produce an unbiased estimate. A more general definition for an unbiased estimator is given by the following:

$$\boxed{E\left\{\hat{\mathbf{x}}_k(\tilde{\mathbf{y}})\right\} = \mathbf{x} \quad \text{for all } k} \tag{2.94}$$

Similar to the batch estimator, it is desired to estimate $\hat{\mathbf{x}}_{k+1}$ as a linear combination of the previous estimate $\hat{\mathbf{x}}_k$ and measurements $\tilde{\mathbf{y}}_{k+1}$ as

$$\hat{\mathbf{x}}_{k+1} = G_{k+1}\hat{\mathbf{x}}_k + K_{k+1}\tilde{\mathbf{y}}_{k+1} \tag{2.95}$$

where G_{k+1} and K_{k+1} are deterministic matrices. To determine the conditions for an unbiased estimator, we begin by assuming that the (sequential) measurement is modelled by

$$\tilde{\mathbf{y}}_{k+1} = H_{k+1}\mathbf{x}_{k+1} + \mathbf{v}_{k+1} \tag{2.96}$$

Substituting eqn. (2.96) into the estimator equation (2.95) gives

$$\hat{\mathbf{x}}_{k+1} = G_{k+1}\hat{\mathbf{x}}_k + K_{k+1}H_{k+1}\mathbf{x}_{k+1} + K_{k+1}\mathbf{v}_{k+1} \tag{2.97}$$

Taking the expectation of both sides of eqn. (2.97) and using eqn. (2.94) gives the following condition for an unbiased estimate:

$$G_{k+1} = I - K_{k+1}H_{k+1} \tag{2.98}$$

[‡] This implies that the estimate is a *function* of the measurements.

Substituting eqn. (2.98) into eqn. (2.95) yields

$$\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k + K_{k+1}(\tilde{\mathbf{y}}_{k+1} - H_{k+1}\hat{\mathbf{x}}_k) \tag{2.99}$$

which clearly has the structure of the sequential estimator in eqn. (1.65). Therefore, the sequential least squares estimator also produces an unbiased estimate. The case for the unbiased estimator with *a priori* estimates is left as an exercise for the reader.

Example 2.1: In this example we will show that the sample variance in eqn. (1.2) produces an unbiased estimate of $\hat{\sigma}^2$. For random data $\{\tilde{y}(t_1), \tilde{y}(t_2), \ldots, \tilde{y}(t_m)\}$ the sample variance is given by

$$\hat{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left[\tilde{y}(t_i) - \hat{\mu}\right]^2$$

For any random variable z, the variance is given by $\text{var}\{z\} = E\{z^2\} - E\{z\}^2$, which is derived from the parallel axis theorem. Defining $E\{\hat{\sigma}^2\} \equiv S^2$, and applying this to the sample variance equation with the definition of the sample mean gives

$$\begin{aligned}
S^2 &= \frac{1}{m-1}\left[\sum_{i=1}^{m} E\left\{[\tilde{y}(t_i)]^2\right\} - \frac{1}{m}E\left\{\left[\sum_{i=1}^{m}\tilde{y}(t_i)\right]^2\right\}\right] \\
&= \frac{1}{m-1}\left[\sum_{i=1}^{m}\left(\sigma^2+\mu^2\right) - \frac{1}{m}\left\{\text{var}\left[\sum_{i=1}^{m}\tilde{y}(t_i)\right] + \left[E\left\{\sum_{i=1}^{m}\tilde{y}(t_i)\right\}\right]^2\right\}\right] \\
&= \frac{1}{m-1}\left[m\sigma^2 + m\mu^2 - \frac{1}{m}m\sigma^2 - \frac{1}{m}m^2\mu^2\right] \\
&= \frac{1}{m-1}\left[m\sigma^2 - \sigma^2\right] \\
&= \sigma^2
\end{aligned}$$

Therefore, this estimator is unbiased. However, the sample variance shown in this example does not give an estimate with the smallest mean-square-error for Gaussian (normal) distributions.[1]

2.3 Maximum Likelihood Estimation

We have seen that minimum variance estimation provides a powerful method to determine least squares estimates through rigorous proof of the relationship between the weight matrix and measurement-error covariance matrix. In this section another powerful method, known as *maximum likelihood estimation* is shown. This method was first introduced by R.A. Fisher, a geneticist and statistician in the 1920s. Maximum likelihood yields estimates for the unknown quantities which maximize the probability of obtaining the observed set of data. Although fundamentally different than minimum variance we will show that under the assumption of zero-mean Gaussian noise measurement-error process, both maximum likelihood and minimum variance estimation yield the same exact results for the least squares estimates.

We begin the topic of maximum likelihood by first considering a probability density function (see Appendix B) which is a function of the measurements and unknown parameters, denoted by $f(\tilde{\mathbf{y}};\mathbf{x})$. For motivational purposes, let $\tilde{\mathbf{y}}$ be a random sample from a simple Gaussian distribution. The density function is given by (see Appendix B)

$$f(\tilde{\mathbf{y}};\mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{m/2} e^{\left[-\sum_{i=1}^{m}(\tilde{y}_i-\mu)^2\Big/\left(2\sigma^2\right)\right]} \tag{2.100}$$

Clearly, the Gaussian distribution is a monotonic exponential function for the mean (μ) and variance (σ^2). Due to the monotonic aspect of the function, this fit can be accomplished by also taking the natural logarithm of eqn. (2.100), which yields

$$\ln\left[f(\tilde{\mathbf{y}};\mathbf{x})\right] = -\frac{m}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(\tilde{y}_i-\mu)^2 \tag{2.101}$$

Now the fit leads immediately to an equivalent quadratic optimization problem to maximize the function in eqn. (2.101). This leads to the concept of maximum likelihood estimation, which is stated as follows. Given a measurement $\tilde{\mathbf{y}}$, the maximum-likelihood estimate $\hat{\mathbf{x}}$ is the value of $\mathbf{x}$ which maximizes $f(\tilde{\mathbf{y}};\mathbf{x})$, which is the likelihood that $\mathbf{x}$ resulted in the measured $\tilde{\mathbf{y}}$.

The *likelihood function* $L(\tilde{\mathbf{y}};\mathbf{x})$ is also a probability density function, given by

$$L(\tilde{\mathbf{y}};\mathbf{x}) = \prod_{i=1}^{p} f_i(\tilde{\mathbf{y}};\mathbf{x}) \tag{2.102}$$

where p is the total number of density functions (a product of a number of density functions, known as a joint density, is also a density function in itself). The goal of the method of maximum likelihood is to choose as our estimate of the unknown parameters $\mathbf{x}$ that value for which the *probability* of obtaining the observations $\tilde{\mathbf{y}}$ is maximized. Many likelihood functions contain exponential terms, which can complicate the mathematics involved in obtaining a solution. However, since $\ln\left[L(\tilde{\mathbf{y}};\mathbf{x})\right]$ is a monotonic function of $L(\tilde{\mathbf{y}};\mathbf{x})$, finding $\mathbf{x}$ to maximize $\ln\left[L(\tilde{\mathbf{y}};\mathbf{x})\right]$ is equivalent to maximizing $L(\tilde{\mathbf{y}};\mathbf{x})$.[§] It follows that for a maximum we have the following:

[§] Also, taking the natural logarithm changes a product to a sum which often simplifies the problem to be solved.

necessary condition

$$\boxed{\left\{\frac{\partial}{\partial \mathbf{x}} \ln\left[L(\tilde{\mathbf{y}};\mathbf{x})\right]\right\}\Bigg|_{\hat{\mathbf{x}}} = \mathbf{0}} \tag{2.103}$$

sufficient condition

$$\frac{\partial^2}{\partial \mathbf{x}\, \partial \mathbf{x}^T} \ln\left[L(\tilde{\mathbf{y}};\mathbf{x})\right] \text{ must be negative definite.} \tag{2.104}$$

Equation (2.103) is often called the *likelihood equation*.[5, 6] Let us demonstrate this method by a few simple examples.

Example 2.2: Let $\tilde{\mathbf{y}}$ be a random sample from a Gaussian distribution. We desire to determine estimates for the mean (μ) and variance (σ^2), so that $\mathbf{x}^T = \left[\mu\ \sigma^2\right]^T$. For this case the likelihood function is given by eqn. (2.100):

$$L(\tilde{\mathbf{y}};\mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{m/2} e^{\left[-\sum\limits_{i=1}^{m}(\tilde{y}_i-\mu)^2 \Big/ (2\sigma^2)\right]}$$

The log likelihood function is given by

$$\ln\left[L(\tilde{\mathbf{y}};\mathbf{x})\right] = -\frac{m}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(\tilde{y}_i-\mu)^2$$

To determine the maximizing μ we take the partial derivative of $\ln\left[L(\tilde{\mathbf{y}};\mathbf{x})\right]$ with respect to μ, evaluated at $\hat{\mu}$, and equate the resultant to zero, giving

$$\left\{\frac{\partial}{\partial \mu}\ln\left[L(\tilde{\mathbf{y}};\hat{\mathbf{x}})\right]\right\}\Bigg|_{\hat{\mu}} = \frac{1}{\sigma^2}\sum_{i=1}^{m}(\tilde{y}_i-\hat{\mu}) = 0$$

Solving for $\hat{\mu}$ yields

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}\tilde{y}_i$$

which is the well known sample mean. To determine the maximizing σ^2 we take the partial derivative of $\ln\left[L(\tilde{\mathbf{y}};\hat{\mathbf{x}})\right]$ with respect to σ^2, evaluated at $\hat{\sigma}^2$, and equate the resultant to zero, giving

$$\left\{\frac{\partial}{\partial \sigma^2}\ln\left[L(\tilde{\mathbf{y}};\hat{\mathbf{x}})\right]\right\}\Bigg|_{\hat{\sigma}^2} = -\frac{m}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{m}(\tilde{y}_i-\mu)^2 = 0$$

Solving for $\hat{\sigma}^2$ yields

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m}(\tilde{y}_i-\mu)^2$$

which is the sample variance. It is easy to show that this estimate for σ^2 is biased, whereas the estimate shown in example 2.1 is unbiased. Thus, two different principles of estimation (unbiased estimator and maximum likelihood) give two different estimators.

Example 2.3: An advantage of using maximum likelihood is that we are not limited to Gaussian distributions. For example, suppose we wish to determine the probability of obtaining a certain number of heads in multiple flips of a coin. We are given $\tilde{y}$ "successes" in n trials, and wish to estimate the probability of success x of the *binomial* distribution.[7] The likelihood function is given by

$$L(\tilde{y};x) = \binom{n}{\tilde{y}} x^{\tilde{y}}(1-x)^{n-\tilde{y}}$$

The log likelihood function is given by

$$\ln\left[L(\tilde{y};x)\right] = \ln\binom{n}{\tilde{y}} + \tilde{y}\ln(x) + (n-\tilde{y})\ln(1-x)$$

To determine the maximizing x we take the partial derivative of $\ln\left[L(\tilde{y};x)\right]$ with respect to x, evaluated at $\hat{x}$, and equate the resultant to zero, giving

$$\left\{\frac{\partial}{\partial x}\ln\left[L(\tilde{y};x)\right]\right\}\bigg|_{\hat{x}} = \frac{\tilde{y}}{\hat{x}} - \frac{n-\tilde{y}}{1-\hat{x}} = 0$$

Therefore, the likelihood function has a maximum at

$$\hat{x} = \frac{\tilde{y}}{n}$$

This intuitively makes sense for our coin toss example, since we expect to obtain a probability of $1/2$ in n flips (for a balanced coin).

Maximum likelihood has many desirable properties. The first is the *invariance principle*,[5] which is stated as follows: Let $\hat{\mathbf{x}}$ be the maximum likelihood estimate of $\mathbf{x}$. Then the maximum likelihood estimate of any function $g(\mathbf{x})$ of these parameters is the function $g(\hat{\mathbf{x}})$ of the maximum likelihood estimate. This is a powerful tool since we do not have to take more partial derivatives to determine the maximum likelihood estimate! A simple example involves estimating the standard deviation, σ, in example 2.2. Using the invariance principle the solution is simply given by $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$. Another property is that maximum likelihood is an *asymptotically efficient* estimator. This means that if the sample size m is large, the maximum likelihood estimate is approximately unbiased and has a variance that approaches the smallest that can be achieved by any estimator.[4] We see that this property is true in example 2.2, since as m becomes large the maximum likelihood estimate for the variance approaches the unbiased estimate asymptotically. Finally, the estimation errors in the maximum likelihood estimate can be shown to be *asymptotically Gaussian* no matter what density function is used in the likelihood function. Proofs of these properties can be found in Sorenson.[5]

We now turn our attention to the least squares problem. We will again use the linear observation model from eqn. (2.1), but we assume that $\mathbf{v}$ has a zero mean Gaussian distribution with covariance given by eqn. (2.19). To compute the maximum likelihood estimate of $\mathbf{x}$ we need the probability density function of $\tilde{\mathbf{y}}$, which we know is Gaussian since measurements of a linear system, such as eqn. (2.1), driven by Gaussian noise are also Gaussian (see Appendix B). To determine the mean of the observation model, we take the expectation of both sides of eqn. (2.1)

$$\boldsymbol{\mu} \equiv E\left\{\tilde{\mathbf{y}}\right\} = E\left\{H\mathbf{x}\right\} + E\left\{\mathbf{v}\right\} \tag{2.105}$$

Since both H and $\mathbf{x}$ are *deterministic* quantities and since $\mathbf{v}$ has zero mean (so that $E\{\mathbf{v}\} = \mathbf{0}$), eqn. (2.105) reduces to

$$\boldsymbol{\mu} = H\mathbf{x} \tag{2.106}$$

Next, we determine the covariance of the observation model, which is given by

$$\text{cov}\left\{\tilde{\mathbf{y}}\right\} \equiv E\left\{\left(\tilde{\mathbf{y}} - \boldsymbol{\mu}\right)\left(\tilde{\mathbf{y}} - \boldsymbol{\mu}\right)^T\right\} \tag{2.107}$$

Substituting eqns. (2.1) and (2.106) into (2.107) gives

$$\text{cov}\left\{\tilde{\mathbf{y}}\right\} = R \tag{2.108}$$

In shorthand notation it is common to use $\tilde{\mathbf{y}} \sim \mathcal{N}(\boldsymbol{\mu}, R)$ to represent a Gaussian (normal) noise process with mean $\boldsymbol{\mu}$ and covariance R. Next, from Appendix B, we use the *multidimensional* or *multivariate normal distribution* for the likelihood function, and from eqns. (2.106) and (2.108) we have

$$L(\tilde{\mathbf{y}};\mathbf{x}) = \frac{1}{(2\pi)^{m/2}\left[\det(R)\right]^{1/2}} \exp\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}} - H\mathbf{x}\right]^T R^{-1}\left[\tilde{\mathbf{y}} - H\mathbf{x}\right]\right\} \tag{2.109}$$

The log likelihood function is given by

$$\ln\left[L(\tilde{\mathbf{y}};\mathbf{x})\right] = -\frac{1}{2}\left[\tilde{\mathbf{y}} - H\mathbf{x}\right]^T R^{-1}\left[\tilde{\mathbf{y}} - H\mathbf{x}\right] - \frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln\left[\det(R)\right] \tag{2.110}$$

We can ignore the last two terms of the right-hand side of eqn. (2.110) since they are independent of $\mathbf{x}$. Also, if we take the negative of eqn. (2.110), then maximizing the log likelihood function to determine the optimal estimate $\hat{\mathbf{x}}$ is equivalent to *minimizing*

$$J(\hat{\mathbf{x}}) = \frac{1}{2}\left[\tilde{\mathbf{y}} - H\hat{\mathbf{x}}\right]^T R^{-1}\left[\tilde{\mathbf{y}} - H\hat{\mathbf{x}}\right] \tag{2.111}$$

The optimal estimate for $\mathbf{x}$ found by minimizing eqn. (2.111) is exactly equivalent to the minimum variance solution given in eqn. (2.29)! Therefore, for the case of Gaussian measurement errors the minimum variance and maximum likelihood estimates are identical to the least squares solution with the weight replaced with the inverse measurement-error covariance. The term $\frac{1}{2}$ in the loss function comes directly from maximum likelihood, which also helps simplify the mathematics when taking partials.

Example 2.4: In example 2.2 we estimated the variance using a random measurement sample from a normal distribution. In this example we will expand upon this to estimate the covariance from a multivariate normal distribution given a set of observations:

$$\left\{\tilde{\mathbf{y}}_1,\, \tilde{\mathbf{y}}_2,\, \ldots,\, \tilde{\mathbf{y}}_p\right\}$$

The likelihood function in this case is the joint density function, given by

$$L(R) = \prod_{i=1}^{p} \frac{1}{(2\pi)^{m/2}\,[\det(R)]^{1/2}} \exp\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]^T R^{-1}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]\right\}$$

The log likelihood function is given by

$$\ln[L(R)] = \sum_{i=1}^{p}\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]^T R^{-1}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right] - \frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln[\det(R)]\right\}$$

To determine an estimate of R we need to take the partial of $\ln[L(R)]$ with respect to R and set the resultant to zero. In order to accomplish this task, we will need to review some matrix calculus differentiating rules. For any given matrices R and G we have

$$\frac{\partial \ln[\det(R)]}{\partial R} = (R^T)^{-1}$$

and

$$\frac{\partial\, \mathrm{Tr}(R^{-1}G)}{\partial R} = -(R^T)^{-1}G(R^T)^{-1}$$

where Tr denotes the trace operator. It can also be shown through simple matrix manipulations that

$$\sum_{i=1}^{p}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]^T R^{-1}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right] = \mathrm{Tr}(R^{-1}G)$$

where

$$G = \sum_{i=1}^{p}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]^T$$

Now, since R is symmetric we have

$$\frac{\partial \ln[L(R)]}{\partial R} = -\frac{p}{2}R^{-1} + \frac{1}{2}R^{-1}GR^{-1}$$

Therefore, the maximum likelihood estimate for the covariance is given by

$$\hat{R} = \frac{1}{p}\sum_{i=1}^{p}\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]\left[\tilde{\mathbf{y}}_i - \boldsymbol{\mu}\right]^T$$

It can also be shown that this estimate is biased.

2.4 Cramér-Rao Inequality

This section describes one of the most useful and important concepts in estimation theory. The Cramér-Rao inequality[8] can be used to give us a lower bound on the expected errors between the estimated quantities and the *true* values from the known statistical properties of the measurement errors. The theory was proved independently by Cramér and Rao, although it was found earlier by Fisher[9] for the special case of a Gaussian distribution. Let $f(\tilde{\mathbf{y}};\mathbf{x})$ be the probability density function of the sample $\tilde{\mathbf{y}}$. The Cramér-Rao inequality for an unbiased estimate $\hat{\mathbf{x}}$ is given by[¶]

$$\boxed{P \equiv E\left\{\left(\hat{\mathbf{x}}-\mathbf{x}\right)\left(\hat{\mathbf{x}}-\mathbf{x}\right)^T\right\} \geq F^{-1}} \tag{2.112}$$

where the *Fisher information matrix*, F, is given by

$$F = E\left\{\left[\frac{\partial}{\partial\mathbf{x}}\ln f(\tilde{\mathbf{y}};\mathbf{x})\right]\left[\frac{\partial}{\partial\mathbf{x}}\ln f(\tilde{\mathbf{y}};\mathbf{x})\right]^T\right\} \tag{2.113}$$

It can be shown that the Fisher information matrix[10] can also be computed using the Hessian matrix, given by

$$F = -E\left\{\frac{\partial^2}{\partial\mathbf{x}\,\partial\mathbf{x}^T}\ln f(\tilde{\mathbf{y}};\mathbf{x})\right\} \tag{2.114}$$

The first- and second-order partial derivatives are assumed to exist and to be absolutely integrable. A formal proof of the Cramér-Rao inequality requires using the "conditions of regularity."[1] However, a slightly different approach is taken here. We begin the proof by using the definition of a probability density function

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\tilde{\mathbf{y}};\mathbf{x})\,d\tilde{y}_1\,d\tilde{y}_2\cdots\,d\tilde{y}_m = 1 \tag{2.115}$$

[¶] For a definition of what it means for one matrix to be greater than another matrix see Appendix A.

In short-hand notation, we write eqn. (2.115) as

$$\int_{-\infty}^{\infty} f(\tilde{\mathbf{y}};\mathbf{x})\, d\tilde{\mathbf{y}} = 1 \tag{2.116}$$

Taking the partial of eqn. (2.116) with respect to $\mathbf{x}$ gives

$$\frac{\partial}{\partial \mathbf{x}} \int_{-\infty}^{\infty} f(\tilde{\mathbf{y}};\mathbf{x})\, d\tilde{\mathbf{y}} = \int_{-\infty}^{\infty} \left[\frac{\partial f(\tilde{\mathbf{y}};\mathbf{x})}{\partial \mathbf{x}}\right]^T d\tilde{\mathbf{y}} = \mathbf{0} \tag{2.117}$$

Next, since $\hat{\mathbf{x}}$ is assumed to be unbiased, we have

$$E\left\{\hat{\mathbf{x}} - \mathbf{x}\right\} = \int_{-\infty}^{\infty} \left(\hat{\mathbf{x}} - \mathbf{x}\right) f\left(\tilde{\mathbf{y}};\mathbf{x}\right) d\tilde{\mathbf{y}} = \mathbf{0} \tag{2.118}$$

Differentiating both sides of eqn. (2.118) with respect to $\mathbf{x}$ gives

$$\int_{-\infty}^{\infty} \left(\hat{\mathbf{x}} - \mathbf{x}\right) \left[\frac{\partial f(\tilde{\mathbf{y}};\mathbf{x})}{\partial \mathbf{x}}\right]^T d\tilde{\mathbf{y}} - I = 0 \tag{2.119}$$

The identity matrix in eqn. (2.119) is obtained since a probability density function always satisfies eqn. (2.116). Next, we use the following logarithmic differentiation rule:[11]

$$\frac{\partial f(\tilde{\mathbf{y}};\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{\partial}{\partial \mathbf{x}} \ln f(\tilde{\mathbf{y}};\mathbf{x})\right] f(\tilde{\mathbf{y}};\mathbf{x}) \tag{2.120}$$

Substituting eqn. (2.120) into eqn. (2.119) leads to

$$I = \int_{-\infty}^{\infty} \left(\mathbf{a}\,\mathbf{b}^T\right) d\tilde{\mathbf{y}} \tag{2.121}$$

where

$$\mathbf{a} \equiv f(\tilde{\mathbf{y}};\mathbf{x})^{1/2} \left(\hat{\mathbf{x}} - \mathbf{x}\right) \tag{2.122a}$$

$$\mathbf{b} \equiv f(\tilde{\mathbf{y}};\mathbf{x})^{1/2} \left[\frac{\partial}{\partial \mathbf{x}} \ln f(\tilde{\mathbf{y}};\mathbf{x})\right] \tag{2.122b}$$

The error-covariance expression in eqn. (2.112) can be rewritten using the definition in eqn. (2.122a) by

$$P = \int_{-\infty}^{\infty} \left(\mathbf{a}\,\mathbf{a}^T\right) d\tilde{\mathbf{y}} \tag{2.123}$$

Also, the Fisher information matrix can be rewritten as

$$F = \int_{-\infty}^{\infty} \left(\mathbf{b}\,\mathbf{b}^T\right) d\tilde{\mathbf{y}} \tag{2.124}$$

Now, multiply eqn. (2.121) on the left by an arbitrary row vector $\boldsymbol{\alpha}^T$ and on the right by an arbitrary column vector $\boldsymbol{\beta}$, so that

$$\boldsymbol{\alpha}^T \boldsymbol{\beta} = \int_{-\infty}^{\infty} \boldsymbol{\alpha}^T \left(\mathbf{a}\,\mathbf{b}^T\right) \boldsymbol{\beta}\, d\tilde{\mathbf{y}} \tag{2.125}$$

Next, we make use of the *Schwartz inequality* (see §A.2), which is given by[‖]

$$\left[\int_{-\infty}^{\infty} g\left(\tilde{\mathbf{y}};\mathbf{x}\right) h\left(\tilde{\mathbf{y}};\mathbf{x}\right) d\tilde{\mathbf{y}}\right]^2 \le \int_{-\infty}^{\infty} g^2\left(\tilde{\mathbf{y}};\mathbf{x}\right) d\tilde{\mathbf{y}} \int_{-\infty}^{\infty} h^2\left(\tilde{\mathbf{y}};\mathbf{x}\right) d\tilde{\mathbf{y}} \tag{2.126}$$

If we let $g\left(\tilde{\mathbf{y}};\mathbf{x}\right) = \boldsymbol{\alpha}^T\mathbf{a}$ and $h\left(\tilde{\mathbf{y}};\mathbf{x}\right) = \mathbf{b}^T\boldsymbol{\beta}$, then eqn. (2.126) becomes

$$\left[\int_{-\infty}^{\infty} \boldsymbol{\alpha}^T(\mathbf{a}\mathbf{b}^T)\boldsymbol{\beta}\, d\tilde{\mathbf{y}}\right]^2 \le \int_{-\infty}^{\infty} \boldsymbol{\alpha}^T(\mathbf{a}\,\mathbf{a}^T)\boldsymbol{\alpha}\, d\tilde{\mathbf{y}} \int_{-\infty}^{\infty} \boldsymbol{\beta}^T(\mathbf{b}\,\mathbf{b}^T)\boldsymbol{\beta}\, d\tilde{\mathbf{y}} \tag{2.127}$$

Using the definitions in eqns. (2.123) and (2.124), and assuming that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are independent of $\tilde{\mathbf{y}}$ gives

$$\left(\boldsymbol{\alpha}^T\boldsymbol{\beta}\right)^2 \le \left(\boldsymbol{\alpha}^T P\boldsymbol{\alpha}\right)\left(\boldsymbol{\beta}^T F\boldsymbol{\beta}\right) \tag{2.128}$$

Finally, choosing the particular choice $\boldsymbol{\beta} = F^{-1}\boldsymbol{\alpha}$ gives

$$\boldsymbol{\alpha}^T(P - F^{-1})\boldsymbol{\alpha} \ge 0 \tag{2.129}$$

Since $\boldsymbol{\alpha}$ is arbitrary then $P \ge F^{-1}$, which proves the Cramér-Rao inequality.

The Cramér-Rao inequality gives a *lower* bound on the expected errors. When the equality in eqn. (2.112) is satisfied, then the estimator is said to be *efficient*. This can be useful for the investigation of the quality of a particular estimator. Therefore, the Cramér-Rao inequality is certainly useful information! It should be stressed that the Cramér-Rao inequality gives a lower bound on the expected errors only for the case of unbiased estimates.

Let us now turn our attention to the Gauss-Markov Theorem in eqn. (2.29). The negative of the log likelihood function for this system is given by $J(\mathbf{x})$ plus terms independent of $\mathbf{x}$. Therefore, to compute the Fisher information matrix we can directly use the loss function in eqn. (2.111) so that

$$F = \frac{\partial^2}{\partial\mathbf{x}\,\partial\mathbf{x}^T}\, J(\mathbf{x}) \tag{2.130}$$

An expectation is not required in eqn. (2.130) since the loss function is quadratic in $\mathbf{x}$. Carrying out this computation leads to

$$F = (H^T R^{-1} H) \tag{2.131}$$

Hence, the Cramér-Rao inequality is given by

$$P \ge (H^T R^{-1} H)^{-1} \tag{2.132}$$

Let us now find an expression for the estimate covariance P. Using eqns. (2.29) and (2.1) leads to

$$\hat{\mathbf{x}} - \mathbf{x} = (H^T R^{-1} H)^{-1} H^T R^{-1}\mathbf{v} \tag{2.133}$$

[‖] If $\int_{-\infty}^{\infty} a(\mathbf{x})b(\mathbf{x})\, d\mathbf{x} = 1$ then $\int_{-\infty}^{\infty} a^2(\mathbf{x})\, d\mathbf{x} \int_{-\infty}^{\infty} b^2(\mathbf{x})\, d\mathbf{x} \ge 1$; the equality holds if $a(\mathbf{x}) = c\, b(\mathbf{x})$ where c is not a function of $\mathbf{x}$.

Using $E\{\mathbf{v}\mathbf{v}^T\} = R$ leads to the following estimate covariance:

$$P = (H^T R^{-1} H)^{-1} \tag{2.134}$$

Therefore, the *equality* in eqn. (2.132) is satisfied, so, the least squares estimate from the Gauss-Markov Theorem is the most efficient possible estimate!

Example 2.5: In this example we will show how the covariance expression in eqn. (2.134) can be used to provide boundaries on the expected errors. For this example a set of 1001 measurement points sampled at 0.01-second intervals was taken using the following observation model:

$$y(t) = \cos(t) + 2\sin(t) + \cos(2t) + 2\sin(3t) + v(t)$$

where $v(t)$ is a zero-mean Gaussian noise process with variance given by $R = 0.01$. The least squares estimator from eqn. (2.29) was used to estimate the coefficients of the transcendental functions. In this example the basis functions used in the estimator are equivalent to the functions in the observation model. Estimates were found from 1000 trial runs using a different random number seed between runs. Statistical conclusions can be made if the least squares solution is performed many times using different measurement sets. This approach is known as *Monte Carlo simulation.* A plot of the actual errors for each estimate and associated 3σ boundaries (found from taking the square root of the diagonal elements of P and multiplying the result by 3) is shown in Figure 2.1. From probability theory, for a Gaussian distribution, there is a 0.9974 probability that the estimate error will be inside of the 3σ boundary. We see that the estimate errors in Figure 2.1 agree with this assessment, since for 1000 trial runs we expect about 3 estimates to be outside of the 3σ boundary. This example clearly shows the power of the estimate covariance and Cramér-Rao lower bound. It is important to note that in this example the estimate covariance, P, can be computed *without* any measurement information, since it only depends on H and R. This powerful tool allows one to use probabilistic concepts to compute estimate error boundaries, and subsequently analyze the expected performance in a dynamical system. This is demonstrated further in Chapter 4.

Example 2.6: In this example we will show the usefulness of the Cramér-Rao inequality for parameter estimation. Suppose we wish to estimate a nonlinear appearing parameter, $a > 0$, of the following exponential model:

$$\tilde{y}_k = B\, e^{a\, t_k} + v_k, \quad k = 1, 2 \ldots, m$$

where v_k is a zero-mean Gaussian white-noise process with variance given by σ^2. We can choose to employ nonlinear least squares to iteratively determine the parameter a, given measurements y_k and a known $B > 0$ coefficient. If this approach is taken,

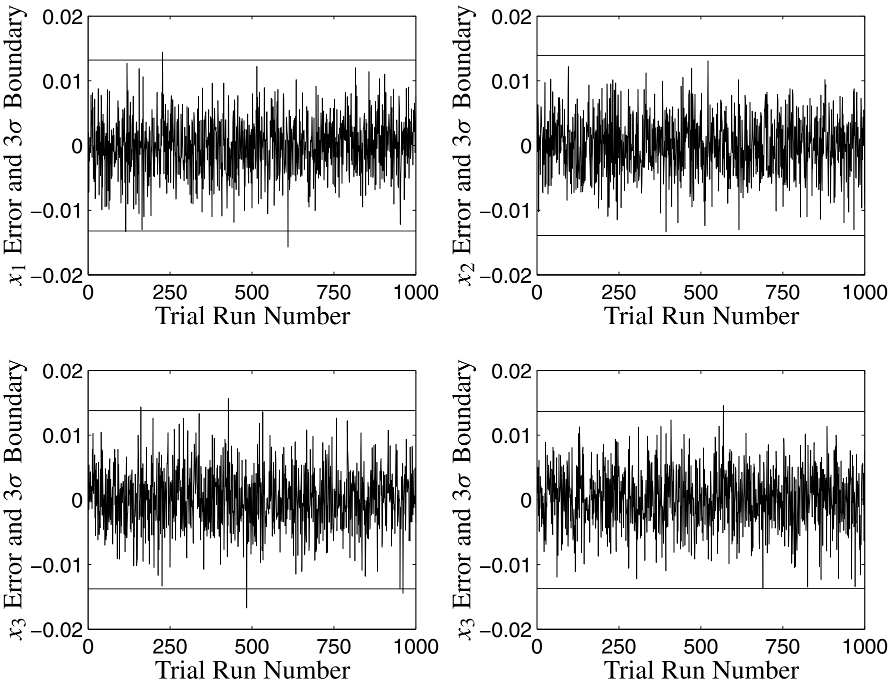

Figure 2.1: Estimate Error and 3σ Boundaries

then the covariance of the estimate error is given by

$$P = \sigma^2 (H^T H)^{-1}$$

where

$$H = \left[B\, t_1\, e^{a\, t_1} \;\; B\, t_2\, e^{a\, t_2} \;\cdots\; B\, t_m\, e^{a\, t_m} \right]^T \tag{2.135}$$

The matrix P is also equivalent to the Cramér-Rao lower bound. Suppose instead we wish to simplify the estimation process by defining $\tilde{z}_k \equiv \ln \tilde{y}_k$, using the change of variables approach shown in Table 1.1. Then, linear squares can be applied to determine a. But how optimal is this solution? It is desired to study the effects of applying this linear approach because the logarithmic function also affects the Gaussian noise. Expanding $\tilde{z}_k$ in a first-order series gives

$$\ln \tilde{y}_k - \ln B \approx a\, t_k + \frac{2\, v_k}{2\, B\, e^{a\, t_k} + v_k}$$

The linear least squares "H matrix," denoted by $\mathcal{H}$, is now simply given by

$$\mathcal{H} = \left[t_1 \; t_2 \; \cdots \; t_m \right]^T \tag{2.136}$$

However, the new measurement noise will certainly not be Gaussian anymore. A first-order expansion of the new measurement noise is given by

$$\varepsilon_k \equiv 2\,v_k(2\,B\,e^{a\,t_k}+v_k)^{-1} \approx \frac{v_k}{B\,e^{a\,t_k}}\left(1-\frac{v_k}{2\,B\,e^{a\,t_k}}\right)$$

The variance of ε_k, denoted by ς_k^2, is derived from

$$\begin{aligned}\varsigma_k^2 &= E\{\varepsilon_k^2\}-E\{\varepsilon_k\}^2\\ &= E\left\{\left(\frac{v_k}{B\,e^{a\,t_k}}-\frac{v_k^2}{2\,B^2e^{2\,a\,t_k}}\right)^2\right\}-\frac{\sigma^4}{4\,B^2e^{4\,a\,t_k}}\end{aligned}$$

This leads to (which is left as an exercise for the reader)

$$\varsigma_k^2 = \frac{\sigma^2}{B^2e^{2\,a\,t_k}}+\frac{\sigma^4}{2\,B^4e^{4\,a\,t_k}}$$

Note that ε_k contains both Gaussian and χ^2 components (see Appendix B). Therefore, the covariance of the linear approach, denoted by $\mathcal{P}$, is given by

$$\mathcal{P} = \left(\mathcal{H}^T \text{diag}\left[\varsigma_1^{-2}\ \varsigma_2^{-2}\ \cdots\ \varsigma_m^{-2}\right]\mathcal{H}\right)^{-1}$$

Notice that $\mathcal{P}$ is equivalent to P if $\sigma^4/(2\,B^4e^{4\,a\,t_k})$ is negligible. If this is not the case, then the Cramér-Rao lower bound is not achieved and the linear approach does not lead to an efficient estimator. This clearly shows how the Cramér-Rao inequality can be particularly useful to help quantify the errors introduced by using an approximate solution instead of the optimal approach. A more practical application of the usefulness of the Cramér-Rao lower bound is given in Ref. [12] and exercise 4.15.

2.5 Nonuniqueness of the Weight Matrix

Here we study the truth that more than one weight matrix in the normal equations can yield identical **x** estimates. Actually two classes of weight matrices (which preserve $\hat{\mathbf{x}}$) exist; the first is rather well known, the second is less known and its implications are more subtle.

We first consider the class of weight matrices which is formed by multiplying all elements of W by some scalar α as

$$W' = \alpha W \tag{2.137}$$

The **x** estimate corresponding to W' follows from eqn. (1.30) as

$$\hat{\mathbf{x}}' = \frac{1}{\alpha}(H^T W H)^{-1} H^T (\alpha W)\tilde{\mathbf{y}} = (H^T W H)^{-1} H^T W \tilde{\mathbf{y}} \tag{2.138}$$

so that

$$\hat{\mathbf{x}}' \equiv \hat{\mathbf{x}} \tag{2.139}$$

Therefore, scaling all elements of W does not (formally) affect the estimate solution $\hat{\mathbf{x}}$. Numerically, possible significant errors may result if extremely small or extremely large values of α are used, due to computed truncation errors.

We now consider a second class of weight matrices obtained by adding a nonzero $(m \times m)$ matrix ΔW to W as

$$W'' = W + \Delta W \tag{2.140}$$

Then the estimate solution $\hat{\mathbf{x}}''$ corresponding to W'' is obtained from eqn. (1.30) as

$$\hat{\mathbf{x}}'' = (H^T W'' H)^{-1} H^T W'' \tilde{\mathbf{y}} \tag{2.141}$$

Substituting eqn. (2.140) into eqn. (2.141) yields

$$\hat{\mathbf{x}}'' = \left[H^T W H + (H^T \Delta W) H\right]^{-1} \left[H^T W \tilde{\mathbf{y}} + (H^T \Delta W)\tilde{\mathbf{y}}\right] \tag{2.142}$$

If $\Delta W \neq 0$ exists such that

$$H^T \Delta W = 0 \tag{2.143}$$

then eqn. (2.142) clearly reduces to

$$\hat{\mathbf{x}}'' = (H^T W H)^{-1} H^T W \tilde{\mathbf{y}} \equiv \hat{\mathbf{x}} \tag{2.144}$$

There are, in fact, an infinity of matrices ΔW satisfying the *orthogonality constraint* in eqn. (2.143). To see this, assume that all elements of ΔW except those in the first column are zero, then eqn. (2.143) becomes

$$H^T \Delta W = \begin{bmatrix} h_{11} & h_{21} & \cdots & h_{m1} \\ h_{12} & h_{22} & \cdots & h_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ h_{1n} & h_{2n} & \cdots & h_{mn} \end{bmatrix} \begin{bmatrix} \Delta W_{11} & 0 & \cdots & 0 \\ \Delta W_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Delta W_{m1} & 0 & \cdots & 0 \end{bmatrix} = 0 \tag{2.145}$$

which yields the scalar equations

$$h_{1i}\Delta W_{11} + h_{2i}\Delta W_{21} + \ldots + h_{mi}\Delta W_{m1} = 0, \quad i = 1, 2, \ldots, n \tag{2.146}$$

Equation (2.146) provides n equations to be satisfied by the m unspecified ΔW_{j1}'s. Since any n of the ΔW_{j1}'s can be determined to satisfy eqns. (2.146), while the remaining $(m-n)$ ΔW_{j1}'s can be given arbitrary values, it follows that an infinity of ΔW matrices satisfy eqn. (2.145) and therefore eqn. (2.143).

The fact that more than one weight matrix yields the same estimates for $\mathbf{x}$ is no cause for alarm though. Interpreting the covariance matrix as the inverse of the measurement-error covariance matrix associated with a specific $\tilde{\mathbf{y}}$ of measurements, the above results imply that one can obtain the same $\mathbf{x}$-estimate from the given measured $\mathbf{y}$-values, for a variety of measurement weights, according to eqn. (2.137) or eqns. (2.140) and (2.143). A most interesting question can be asked regarding the covariance matrix of the estimated parameters. From eqn. (2.134), we established that the estimate covariance is

$$P = (H^T W H)^{-1}, \quad W = R^{-1} \tag{2.147}$$

For the first class of weight matrices $W' = \alpha W$ note that

$$P' = \frac{1}{\alpha}(H^T W H)^{-1} = \frac{1}{\alpha}(H^T R^{-1} H)^{-1} \tag{2.148}$$

or

$$P' = \frac{1}{\alpha} P \tag{2.149}$$

Thus linear scaling of the observation weight matrix results in reciprocal linear scaling of the estimate covariance matrix, an intuitively reasonable result.

Considering now the second class of error covariance matrices $W'' = W + \Delta W$, with $H^T \Delta W = 0$, it follows from eqn. (2.147) that

$$P'' = (H^T W H + H^T \Delta W H)^{-1} = (H^T W H)^{-1} \tag{2.150}$$

or

$$P'' = P \tag{2.151}$$

Thus, the additive class of observation weight matrices preserves not only the $\mathbf{x}$-estimates, but also the associated estimate covariance matrix. It may prove possible, in some applications, to exploit this truth since a family of measurement-error covariances can result in the same estimates and associated uncertainties.

Example 2.7: Given the following linear system:

$$\tilde{\mathbf{y}} = H\mathbf{x}$$

with

$$\tilde{\mathbf{y}} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 3 \\ 2 & 2 \\ 3 & 4 \end{bmatrix}$$

For each of the three weight matrices

$$W = I, \quad W' = 3W, \quad W'' = W + \begin{bmatrix} 1/4 & 5/8 & -1/2 \\ 5/8 & 25/16 & -5/4 \\ -1/2 & -5/4 & 1 \end{bmatrix}$$

determine the least squares estimates

$$\hat{\mathbf{x}} = (H^T W H)^{-1} H^T W \tilde{\mathbf{y}}$$
$$\hat{\mathbf{x}}' = (H^T W' H)^{-1} H^T W' \tilde{\mathbf{y}}$$
$$\hat{\mathbf{x}}'' = (H^T W'' H)^{-1} H^T W'' \tilde{\mathbf{y}}$$

and corresponding error-covariance matrices

$$P = (H^T W H)^{-1}$$
$$P' = (H^T W' H)^{-1}$$
$$P'' = (H^T W'' H)^{-1}$$

The reader can verify the numerical results

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}' = \hat{\mathbf{x}}'' = \begin{bmatrix} -1/15 \\ 11/15 \end{bmatrix}$$

and

$$P = P'' = \begin{bmatrix} 29/45 & -19/45 \\ -19/45 & 14/45 \end{bmatrix}$$
$$P' = \frac{1}{3} P = \begin{bmatrix} 29/135 & -19/135 \\ -19/135 & 14/135 \end{bmatrix}$$

These results are consistent with eqns. (2.139), (2.144), (2.149), and (2.151).

2.6 Bayesian Estimation

The parameters that we have estimated in this chapter have been assumed to be unknown constants. In Bayesian estimation, we consider that these parameters are random variables with some *a priori* distribution. Bayesian estimation combines this *a priori* information with the measurements through a conditional density function of $\mathbf{x}$ given the measurements $\tilde{\mathbf{y}}$. This conditional function is known as the *a posteriori distribution* of $\mathbf{x}$. Therefore, Bayesian estimation requires the probability density functions of both the measurement noise and unknown parameters. The posterior density function $f(\mathbf{x}|\tilde{\mathbf{y}})$ for $\mathbf{x}$ (taking the measurements $\tilde{\mathbf{y}}$ into account) is given by *Bayes rule* (see Appendix B for details):

$$\boxed{f(\mathbf{x}|\tilde{\mathbf{y}}) = \frac{f(\tilde{\mathbf{y}}|\mathbf{x})\, f(\mathbf{x})}{f(\tilde{\mathbf{y}})}} \tag{2.152}$$

Note since $\tilde{\mathbf{y}}$ is treated as a set of known quantities, then $f(\tilde{\mathbf{y}})$ is just a normalization factor to ensure that $f(\mathbf{x}|\tilde{\mathbf{y}})$ is a probability density function. Therefore,

$$f(\tilde{\mathbf{y}}) = \int_{-\infty}^{\infty} f(\tilde{\mathbf{y}}|\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x} \tag{2.153}$$

If the integral in eqn. (2.153) exists then the posterior function $f(\mathbf{x}|\tilde{\mathbf{y}})$ is said to be *proper*; if it does not exist then $f(\mathbf{x}|\tilde{\mathbf{y}})$ is *improper*, in which case we let $f(\tilde{\mathbf{y}}) = 1$ (see [13] for sufficient conditions).

Maximum *a posteriori* (MAP) estimation finds an estimate for $\mathbf{x}$ that maximizes eqn. (2.152).[6] Since $f(\tilde{\mathbf{y}})$ does not depend on $\mathbf{x}$, this is equivalent to maximizing $f(\tilde{\mathbf{y}}|\mathbf{x})\, f(\mathbf{x})$. We can again use the natural logarithm (as shown in §2.3) to simplify the problem by maximizing

$$J_{\text{MAP}}(\hat{\mathbf{x}}) = \ln\left[f(\tilde{\mathbf{y}}|\hat{\mathbf{x}})\right] + \ln\left[f(\hat{\mathbf{x}})\right] \tag{2.154}$$

The first term in the sum is actually the log-likelihood function, and the second term gives the *a priori* information on the to-be-determined parameters. Therefore, the MAP estimator maximizes

$$\boxed{J_{\text{MAP}}(\hat{\mathbf{x}}) = \ln\left[L(\tilde{\mathbf{y}}|\hat{\mathbf{x}})\right] + \ln\left[f(\hat{\mathbf{x}})\right]} \tag{2.155}$$

Maximum *a posteriori* estimation has the following properties: (1) if the *a priori* distribution $f(\hat{\mathbf{x}})$ is uniform, then MAP estimation is equivalent to maximum likelihood estimation, (2) MAP estimation shares the asymptotic consistency and efficiency properties of maximum likelihood estimation, (3) the MAP estimator converges to the maximum likelihood estimator for large samples, and (4) the MAP estimator also obeys the invariance principle.

Example 2.8: Suppose we wish to estimate the mean μ of a Gaussian variable from a sample of m independent measurements known to have a standard deviation of $\sigma_{\tilde{y}}$. We have been given that the *a priori* density function of μ is also Gaussian with zero mean and standard deviation σ_μ. The density functions are therefore given by

$$f(\tilde{y}_i|\mu) = \frac{1}{\sigma_{\tilde{y}}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(\tilde{y}_i-\mu)^2}{\sigma_{\tilde{y}}^2}\right\}, \quad i = 1, 2, \ldots, m$$

and

$$f(\mu) = \frac{1}{\sigma_\mu\sqrt{2\pi}} \exp\left\{-\frac{\mu^2}{2\sigma_\mu^2}\right\}$$

Since the measurements are independent we can write

$$f(\tilde{\mathbf{y}}|\mu) = \frac{1}{(\sigma_{\tilde{y}}\sqrt{2\pi})^m} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\frac{(\tilde{y}_i-\mu)^2}{\sigma_{\tilde{y}}^2}\right\}$$

Using eqn. (2.154) and ignoring terms independent of μ we now seek to maximize

$$J_{\text{MAP}}(\hat{\mu}) = -\frac{1}{2}\left[\sum_{i=1}^{m}\frac{\left(\tilde{y}_i-\hat{\mu}\right)^2}{\sigma_{\tilde{y}}^2}+\frac{\hat{\mu}^2}{\sigma_{\mu}^2}\right]$$

Taking the partial of this equation with respect to $\hat{\mu}$ and equating the resultant to zero gives

$$\sum_{i=1}^{m}\frac{\left(\tilde{y}_i-\hat{\mu}\right)}{\sigma_{\tilde{y}}^2}-\frac{\hat{\mu}}{\sigma_{\mu}^2}=0$$

Recall that the maximum likelihood estimate for the mean from example 2.2 is given by

$$\hat{\mu}_{\text{ML}} = \frac{1}{m}\sum_{i=1}^{m}\tilde{y}_i$$

Therefore we can write the maximum *a posteriori* estimate for the mean as

$$\hat{\mu} = \frac{\sigma_{\mu}^2}{\frac{1}{m}\sigma_{\tilde{y}}^2+\sigma_{\mu}^2}\,\hat{\mu}_{\text{ML}}$$

Notice that $\hat{\mu}\to\hat{\mu}_{\text{ML}}$ as either $\sigma_{\mu}^2\to\infty$ or as $m\to\infty$. This is consistent with the properties discussed previously of a maximum *a posteriori* estimator.

Maximum *a posteriori* estimation can also be used to find an optimal estimator for the case with *a priori* estimates, modelled using eqns. (2.43) through (2.46). The assumed probability density functions for this case are given by

$$L(\tilde{\mathbf{y}}|\hat{\mathbf{x}}) = f(\tilde{\mathbf{y}}|\hat{\mathbf{x}}) = \frac{1}{(2\pi)^{m/2}\left[\det(R)\right]^{1/2}}\exp\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}}-H\hat{\mathbf{x}}\right]^T R^{-1}\left[\tilde{\mathbf{y}}-H\hat{\mathbf{x}}\right]\right\} \tag{2.156}$$

$$f(\hat{\mathbf{x}}) = \frac{1}{(2\pi)^{n/2}\left[\det(Q)\right]^{1/2}}\exp\left\{-\frac{1}{2}\left[\hat{\mathbf{x}}_a-\hat{\mathbf{x}}\right]^T Q^{-1}\left[\hat{\mathbf{x}}_a-\hat{\mathbf{x}}\right]\right\} \tag{2.157}$$

Maximizing eqn. (2.155) leads to the following estimator:

$$\boxed{\hat{\mathbf{x}} = \left(H^T R^{-1} H + Q^{-1}\right)^{-1}\left(H^T R^{-1}\tilde{\mathbf{y}} + Q^{-1}\hat{\mathbf{x}}_a\right)} \tag{2.158}$$

which is the same result obtained through minimum variance. However, the solution using MAP estimation is much simpler since we do not need to solve a constrained minimization problem using Lagrange multipliers.

The Cramér-Rao inequality can be extended for a Bayesian estimator. The Cramér-Rao inequality for the case of *a priori* information is given by[5, 14]

$$\boxed{\begin{aligned} P &\equiv E\left\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T\right\} \\ &\geq \left[F+E\left\{\left[\frac{\partial}{\partial \mathbf{x}}\ln f(\mathbf{x})\right]\left[\frac{\partial}{\partial \mathbf{x}}\ln f(\mathbf{x})\right]^T\right\}\right]^{-1} \end{aligned}} \tag{2.159}$$

This can be used to test the efficiency of the MAP estimator. The Fisher information matrix has been computed in eqn. (2.131) as

$$F = \left(H^T R^{-1} H\right) \tag{2.160}$$

Using the *a priori* density function in eqn. (2.157) leads to

$$\begin{aligned} E\left\{\left[\frac{\partial}{\partial \mathbf{x}}\ln f(\mathbf{x})\right]\left[\frac{\partial}{\partial \mathbf{x}}\ln f(\mathbf{x})\right]^T\right\} &= Q^{-1}E\left\{(\hat{\mathbf{x}}_a-\mathbf{x})^T(\hat{\mathbf{x}}_a-\mathbf{x})\right\}Q^{-1} \\ &= Q^{-1}E\left\{\mathbf{w}\,\mathbf{w}^T\right\}Q^{-1} = Q^{-1} \end{aligned} \tag{2.161}$$

Next, we need to compute the covariance matrix P. From eqn. (2.81) and using $MH+N=I$, the estimate can be written as

$$\hat{\mathbf{x}} = \mathbf{x} + M\mathbf{v} + N\mathbf{w} \tag{2.162}$$

Using the definitions in eqns. (2.44) and (2.46), and assuming that $E\left\{\mathbf{v}\,\mathbf{w}^T\right\}=0$ and $E\left\{\mathbf{w}\,\mathbf{v}^T\right\}=0$, the covariance matrix can be written as

$$P = MRM^T + NQN^T \tag{2.163}$$

From the solutions for M and N in eqns. (2.77) and (2.78), the covariance matrix becomes

$$P = \left(H^T R^{-1} H + Q^{-1}\right)^{-1} \tag{2.164}$$

Therefore, the lower bound in the Cramér-Rao inequality is achieved, and thus the estimator (2.158) is efficient. Equation (2.164) can be alternatively written using the matrix inversion lemma, shown by eqns. (1.69) and (1.70), as

$$P = Q - QH^T\left(R + HQH^T\right)^{-1}HQ \tag{2.165}$$

Equation (2.165) may be preferred over eqn. (2.164) if the dimension of R is less than the dimension of Q.

Another approach for Bayesian estimation is a minimum risk (MR) estimator.[14, 15] In practical engineering problems, we are often faced with making a decision in the face of uncertainty. An example involves finding the best value for an aircraft model parameter given wind tunnel data in the face of measurement error uncertainty. Bayesian estimation chooses a *course of action* that has the largest expectation of gain (or smallest expectation of loss). This approach assumes the existence (or at least a guess) of the *a priori* probability function. Minimum risk estimators also use this information to find the best estimate based on decision theory, which assigns a cost to any loss suffered due to errors in the estimate. Our goal is to evaluate the cost $c(\mathbf{x}^*|\mathbf{x})$ of believing that the value of the estimate is $\mathbf{x}^*$ when it is actually $\mathbf{x}$. Since $\mathbf{x}$ is unknown, the actual cost cannot be evaluated; however, we usually assume that $\mathbf{x}$ is distributed by the *a posteriori* function. This approach minimizes the risk, defined as the mean of the cost over all possible values of $\mathbf{x}$, given a set of observations $\tilde{\mathbf{y}}$. The risk function is given by

$$J_{\text{MR}}(\mathbf{x}^*) = \int_{-\infty}^{\infty} c(\mathbf{x}^*|\mathbf{x})\, f(\mathbf{x}|\tilde{\mathbf{y}})\, d\mathbf{x} \tag{2.166}$$

Using Bayes rule we can rewrite the risk as

$$\boxed{J_{\text{MR}}(\mathbf{x}^*) = \int_{-\infty}^{\infty} c(\mathbf{x}^*|\mathbf{x}) \frac{f(\tilde{\mathbf{y}}|\mathbf{x})\, f(\mathbf{x})}{f(\tilde{\mathbf{y}})}\, d\mathbf{x}} \tag{2.167}$$

The *minimum risk* estimate is defined as the value of $\mathbf{x}^*$ that minimizes the loss function in eqn. (2.167).

A common choice for the cost $c(\mathbf{x}^*|\mathbf{x})$ is a quadratic function taking the form

$$c(\mathbf{x}|\mathbf{x}^*) = \frac{1}{2}(\mathbf{x}^* - \mathbf{x})^T S (\mathbf{x}^* - \mathbf{x}) \tag{2.168}$$

where S is a positive definite weighting matrix. The risk is now given by

$$J_{\text{MR}}(\mathbf{x}^*) = \frac{1}{2}\int_{-\infty}^{\infty} (\mathbf{x}^* - \mathbf{x})^T S (\mathbf{x}^* - \mathbf{x})\, f(\mathbf{x}|\tilde{\mathbf{y}})\, d\mathbf{x} \tag{2.169}$$

To determine the minimum risk estimate we take the partial of eqn. (2.169) with respect to $\mathbf{x}^*$, evaluated at $\hat{\mathbf{x}}$, and set the resultant to zero:

$$\left.\frac{\partial J_{\text{MR}}(\mathbf{x}^*)}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}} = \mathbf{0} = S\int_{-\infty}^{\infty} (\hat{\mathbf{x}} - \mathbf{x})\, f(\mathbf{x}|\tilde{\mathbf{y}})\, d\mathbf{x} \tag{2.170}$$

Since S is invertible eqn. (2.170) simply reduces down to

$$\hat{\mathbf{x}}\int_{-\infty}^{\infty} f(\mathbf{x}|\tilde{\mathbf{y}})\, d\mathbf{x} = \int_{-\infty}^{\infty} \mathbf{x}\, f(\mathbf{x}|\tilde{\mathbf{y}})\, d\mathbf{x} \tag{2.171}$$

The integral on the left-hand side of eqn. (2.171) is clearly unity, so that

$$\hat{\mathbf{x}} = \int_{-\infty}^{\infty} \mathbf{x}\, f(\mathbf{x}|\tilde{\mathbf{y}})\, d\mathbf{x} \equiv E\left\{\mathbf{x}|\tilde{\mathbf{y}}\right\} \tag{2.172}$$

Notice that the minimum risk estimator is independent of S in this case. Additionally, the optimal estimate is seen to be the expected value (i.e., the mean) of $\mathbf{x}$ given the measurements $\tilde{\mathbf{y}}$. From Bayes rule we can rewrite eqn. (2.172) as

$$\hat{\mathbf{x}} = \int_{-\infty}^{\infty} \mathbf{x} \frac{f(\tilde{\mathbf{y}}|\mathbf{x})\, f(\mathbf{x})}{f(\tilde{\mathbf{y}})}\, d\mathbf{x} \tag{2.173}$$

We will now use the minimum risk approach to determine an optimal estimate with *a priori* information. Recall from §2.1.2 that we have the following models:

$$\tilde{\mathbf{y}} = H\mathbf{x} + \mathbf{v} \tag{2.174a}$$

$$\hat{\mathbf{x}}_a = \mathbf{x} + \mathbf{w} \tag{2.174b}$$

with associated known expectations and covariances

$$E\{\mathbf{v}\} = \mathbf{0} \tag{2.175a}$$

$$\text{cov}\{\mathbf{v}\} = E\left\{\mathbf{v}\mathbf{v}^T\right\} = R \tag{2.175b}$$

and

$$E\{\mathbf{w}\} = \mathbf{0} \tag{2.176a}$$

$$\text{cov}\{\mathbf{w}\} = E\left\{\mathbf{w}\mathbf{w}^T\right\} = Q \tag{2.176b}$$

Also, recall that $\mathbf{x}$ is now a random variable with associated expectation and covariance

$$E\{\mathbf{x}\} = \hat{\mathbf{x}}_a \tag{2.177a}$$

$$\text{cov}\{\mathbf{x}\} = E\left\{\mathbf{x}\mathbf{x}^T\right\} - E\{\mathbf{x}\}\, E\{\mathbf{x}\}^T = Q \tag{2.177b}$$

The probability functions for $f(\tilde{\mathbf{y}}, \mathbf{x})$ and $f(\mathbf{x})$ are given by

$$f(\tilde{\mathbf{y}}|\mathbf{x}) = \frac{1}{(2\pi)^{m/2}\,[\det(R)]^{1/2}} \exp\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}} - H\mathbf{x}\right]^T R^{-1} \left[\tilde{\mathbf{y}} - H\mathbf{x}\right]\right\} \tag{2.178}$$

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\,[\det(Q)]^{1/2}} \exp\left\{-\frac{1}{2}\left[\hat{\mathbf{x}}_a - \mathbf{x}\right]^T Q^{-1} \left[\hat{\mathbf{x}}_a - \mathbf{x}\right]\right\} \tag{2.179}$$

We now need to determine the density function $f(\tilde{\mathbf{y}})$. Since a sum of Gaussian random variables is itself a Gaussian random variable, then we know that $f(\tilde{\mathbf{y}})$ must also be Gaussian. The mean of $\tilde{\mathbf{y}}$ is simply

$$E\left\{\tilde{\mathbf{y}}\right\} = E\{H\mathbf{x}\} = H\hat{\mathbf{x}}_a \tag{2.180}$$

Assuming that $\mathbf{x}$, $\mathbf{v}$, and $\mathbf{w}$ are uncorrelated with each other, the covariance of $\tilde{\mathbf{y}}$ is given by

$$\begin{aligned} \text{cov}\left\{\tilde{\mathbf{y}}\right\} &= E\left\{\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\right\} - E\left\{\tilde{\mathbf{y}}\right\} E\left\{\tilde{\mathbf{y}}\right\}^T \\ &= E\left\{H\mathbf{w}\mathbf{w}^T H^T\right\} + E\left\{\mathbf{v}\mathbf{v}^T\right\} \end{aligned} \tag{2.181}$$

Therefore, using eqns. (2.175) and (2.176), then eqn. (2.181) can be written as

$$\text{cov}\left\{\tilde{\mathbf{y}}\right\} = HQH^T + R \equiv D \tag{2.182}$$

Hence, $f(\tilde{\mathbf{y}})$ is given by

$$f(\tilde{\mathbf{y}}) = \frac{1}{(2\pi)^{m/2}\left[\det(D)\right]^{1/2}} \exp\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}} - H\hat{\mathbf{x}}_a\right]^T D^{-1}\left[\tilde{\mathbf{y}} - H\hat{\mathbf{x}}_a\right]\right\} \tag{2.183}$$

Using Bayes rule and the matrix inversion lemma shown by eqns. (1.69) and (1.70), it can be shown that $f(\mathbf{x}|\tilde{\mathbf{y}})$ is given by

$$\begin{aligned} f(\mathbf{x}|\tilde{\mathbf{y}}) = & \frac{\left[\det\left(HQH^T + R\right)\right]^{1/2}}{(2\pi)^{n/2}\left[\det(R)\right]^{1/2}\left[\det(Q)\right]^{1/2}} \\ & \times \exp\left\{-\frac{1}{2}\left[\mathbf{x} - H\mathbf{p}\right]^T \left(H^T R^{-1} H + Q^{-1}\right)\left[\mathbf{x} - H\mathbf{p}\right]\right\} \end{aligned} \tag{2.184}$$

where

$$\mathbf{p} = \left(H^T R^{-1} H + Q^{-1}\right)^{-1}\left(H^T R^{-1}\tilde{\mathbf{y}} + Q^{-1}\hat{\mathbf{x}}_a\right) \tag{2.185}$$

Clearly, since eqn. (2.172) is $E\left\{\mathbf{x}|\tilde{\mathbf{y}}\right\}$, then the minimum risk estimate is given by

$$\hat{\mathbf{x}} = \mathbf{p} = \left(H^T R^{-1} H + Q^{-1}\right)^{-1}\left(H^T R^{-1}\tilde{\mathbf{y}} + Q^{-1}\hat{\mathbf{x}}_a\right) \tag{2.186}$$

which is equivalent to the estimate found by minimum variance and maximum *a posteriori*.

The minimum risk approach can be useful since it incorporates a decision-based means to determine the optimal estimate. However, there are many practical disadvantages. Although an analytical solution for the minimum risk using Gaussian distributions can be found in many cases, the evaluation of the integral in eqn. (2.173) may be impractical for general distributions. Also, the minimum risk estimator does not (in general) converge to the maximum likelihood estimate for uniform *a priori* distributions. Finally, unlike maximum likelihood, the minimum risk estimator is not invariant under reparameterization. For these reasons, minimum risk approaches are often avoided in practical estimation problems.

Some important properties of the *a priori* estimator in eqn. (2.186) are given by the following:

$$E\left\{(\mathbf{x} - \hat{\mathbf{x}})\tilde{\mathbf{y}}^T\right\} = 0 \tag{2.187}$$

$$E\left\{(\mathbf{x} - \hat{\mathbf{x}})\hat{\mathbf{x}}^T\right\} = 0 \tag{2.188}$$

The proof of these relations now follows. We first substitute $\hat{\mathbf{x}}$ from eqn. (2.186) into eqn. (2.187), with use of the model given in eqn. (2.174a). Then taking the expectation of the resultant, with $E\left\{\mathbf{v}\,\mathbf{x}^T\right\} = E\left\{\mathbf{x}\,\mathbf{v}^T\right\} = 0$, and using eqn. (2.177a) gives

$$\begin{aligned} E\left\{(\mathbf{x}-\hat{\mathbf{x}})\tilde{\mathbf{y}}^T\right\} &= (I - K H^T R^{-1} H) E\left\{\mathbf{x}\mathbf{x}^T\right\} H^T \\ &\quad - K Q^{-1}\hat{\mathbf{x}}_a\hat{\mathbf{x}}_a^T H^T - K H^T \end{aligned} \tag{2.189}$$

where

$$K \equiv \left(H^T R^{-1} H + Q^{-1}\right)^{-1} \tag{2.190}$$

Next, using the following identity:

$$(I - K H^T R^{-1} H) = K Q^{-1} \tag{2.191}$$

yields

$$E\left\{(\mathbf{x}-\hat{\mathbf{x}})\tilde{\mathbf{y}}^T\right\} = K\left(Q^{-1} E\left\{\mathbf{x}\mathbf{x}^T\right\} H^T - Q^{-1}\hat{\mathbf{x}}_a\hat{\mathbf{x}}_a^T H^T - H^T\right) \tag{2.192}$$

Finally, using eqn. (2.177b) in eqn. (2.192) leads to

$$E\left\{(\mathbf{x}-\hat{\mathbf{x}})\tilde{\mathbf{y}}^T\right\} = 0 \tag{2.193}$$

To prove eqn. (2.188), we substitute eqn. (2.186) into eqn. (2.188), again with use of the model given in eqn. (2.174a). Taking the appropriate expectations leads to

$$\begin{aligned} E\left\{(\mathbf{x}-\hat{\mathbf{x}})\hat{\mathbf{x}}^T\right\} &= E\left\{\mathbf{x}\mathbf{x}^T\right\} H^T R^{-1} H K + \hat{\mathbf{x}}_a\hat{\mathbf{x}}_a^T Q^{-1} K \\ &\quad - K H^T R^{-1} H E\left\{\mathbf{x}\mathbf{x}^T\right\} H^T R^{-1} H K \\ &\quad - K H^T R^{-1} H\hat{\mathbf{x}}_a\hat{\mathbf{x}}_a^T Q^{-1} K - K H^T R^{-1} H K \\ &\quad - K Q^{-1}\hat{\mathbf{x}}_a\hat{\mathbf{x}}_a^T H^T R^{-1} H K - K Q^{-1}\hat{\mathbf{x}}_a\hat{\mathbf{x}}_a^T Q^{-1} K \end{aligned} \tag{2.194}$$

Next, using eqn. (2.177b) and the identity in eqn. (2.191) leads to

$$E\left\{(\mathbf{x}-\hat{\mathbf{x}})\hat{\mathbf{x}}^T\right\} = 0 \tag{2.195}$$

Equations (2.187) and (2.188) show that the residual error is orthogonal to both the measurements and the estimates. Therefore, the concepts shown in §1.6.4 also apply to the *a priori* estimator.

2.7 Advanced Topics

In this section we will show some advanced topics used in probabilistic estimation. As in Chapter 1 we encourage the interested reader to pursue these topics further in the references provided.

2.7.1 Analysis of Covariance Errors

In §2.5 an analysis was shown for simple errors in the measurement-error covariance matrix. In this section we expand upon these results to the case of general errors in the assumed measurement-error covariance matrix. Say that the assumed measurement-error covariance is denoted by $\tilde{R}$ and the actual covariance is denoted by R. The least squares estimate with the assumed covariance matrix is given by

$$\hat{\mathbf{x}} = (H^T \tilde{R}^{-1} H)^{-1} H^T \tilde{R}^{-1} \tilde{\mathbf{y}} \tag{2.196}$$

Using the measurement model in eqn. (2.1) leads to the following residual error:

$$\hat{\mathbf{x}} - \mathbf{x} = (H^T \tilde{R}^{-1} H)^{-1} H^T \tilde{R}^{-1} \mathbf{v} \tag{2.197}$$

The estimate $\hat{\mathbf{x}}$ is unbiased since $E\{\mathbf{v}\} = \mathbf{0}$. Using $E\left\{\mathbf{v}\,\mathbf{v}^T\right\} = R$, the estimate covariance is given by

$$\tilde{P} = (H^T \tilde{R}^{-1} H)^{-1} H^T \tilde{R}^{-1} R\, \tilde{R}^{-1} H (H^T \tilde{R}^{-1} H)^{-1} \tag{2.198}$$

Clearly $\tilde{P}$ reduces to $(H^T R^{-1} H)^{-1}$ when $\tilde{R} = R$ or when H is square (i.e., $m = n$). Next, we define the following relative inefficiency parameter e, which gives a measure of the error induced by the incorrect measurement-error covariance:

$$\boxed{e = \frac{\det\left[(H^T \tilde{R}^{-1} H)^{-1} H^T \tilde{R}^{-1} R\, \tilde{R}^{-1} H (H^T \tilde{R}^{-1} H)^{-1}\right]}{\det\left[(H^T R^{-1} H)^{-1}\right]}} \tag{2.199}$$

We will now prove that $e \geq 1$. Since for any invertible matrix A, $\det(A^{-1}) = 1\big/\det(A)$, eqn. (2.199) reduces to

$$e = \frac{\det(H^T \tilde{R}^{-1} R\, \tilde{R}^{-1} H) \det(H^T R^{-1} H)}{\det(H^T \tilde{R}^{-1} H)^2} \tag{2.200}$$

Performing a singular value decomposition of the matrix $\tilde{R}^{1/2} H$ gives

$$\tilde{R}^{1/2} H = X\, S\, Y^T \tag{2.201}$$

where X and Y are orthogonal matrices.[16] Also, define the following matrix:

$$D \equiv X^T \tilde{R}^{-1/2} R\, \tilde{R}^{-1/2} X \tag{2.202}$$

Using the definitions in eqns. (2.201) and (2.202), then eqn. (2.200) can be written as

$$e = \frac{\det(Y\, S^T D\, S\, Y^T) \det(Y\, S^T D^{-1} S\, Y^T)}{\det(Y\, S^T S\, Y^T)} \tag{2.203}$$

This can easily be reduced to give

$$e = \frac{\det(S^T D\, S) \det(S^T D^{-1} S)}{\det(S^T S)^2} \tag{2.204}$$

Next, we partition the $m \times n$ matrix S into an $n \times n$ matrix S_1 and an $(m-n) \times n$ matrix of zeros so that

$$S = \begin{bmatrix} S_1 \\ 0 \end{bmatrix} \tag{2.205}$$

where S_1 is a diagonal matrix of the singular values. Also, partition D as

$$D = \begin{bmatrix} D_1 & F \\ F^T & D_2 \end{bmatrix} \tag{2.206}$$

where D_1 is a square matrix with the same dimension as S_1 and D_2 is also square. The inverse of D is given by (see Appendix A)

$$D^{-1} = \begin{bmatrix} (D_1 - F\, D_2^{-1} F^T)^{-1} & G \\ G^T & (D_2 - F^T D_1^{-1} F)^{-1} \end{bmatrix} \tag{2.207}$$

where the closed-form expression for G is not required in this development. Substituting eqns. (2.205), (2.206), and (2.207) into eqn. (2.204) leads to

$$e = \frac{\det(D_1)}{\det(D_1 - F\, D_2^{-1} F^T)} \tag{2.208}$$

Next, we use the following identity (see Appendix A):

$$\det(D) = \det(D_2)\det(D_1 - F\, D_2^{-1} F^T) \tag{2.209}$$

which reduces eqn. (2.208) to

$$e = \frac{\det(D_1)\det(D_2)}{\det(D)} \tag{2.210}$$

By Fischer's inequality[16] $e \geq 1$. The specific value of e gives an indication of the inefficiency of the estimator, and can be used to perform a sensitivity analysis given bounds on matrix R. A larger value for e means that the estimates are further (in a statistical sense) from their true values.

Example 2.9: In this simple example we consider a two measurement case with the true covariance given by the identity matrix. The assumed covariance $\tilde{R}$ and H matrices are given by

$$\tilde{R} = \begin{bmatrix} 1+\alpha & 0 \\ 0 & 1+\beta \end{bmatrix}, \quad H = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

where α and β can vary from -0.99 to 1. A three-dimensional plot of the inefficiency in eqn. (2.199) for varying α and β is shown in Figure 2.2. The minimum value (1) is given when $\alpha = \beta = 0$ as expected. Also, the values for e are significantly lower when both α and β are greater than 1 (the average value for e in this

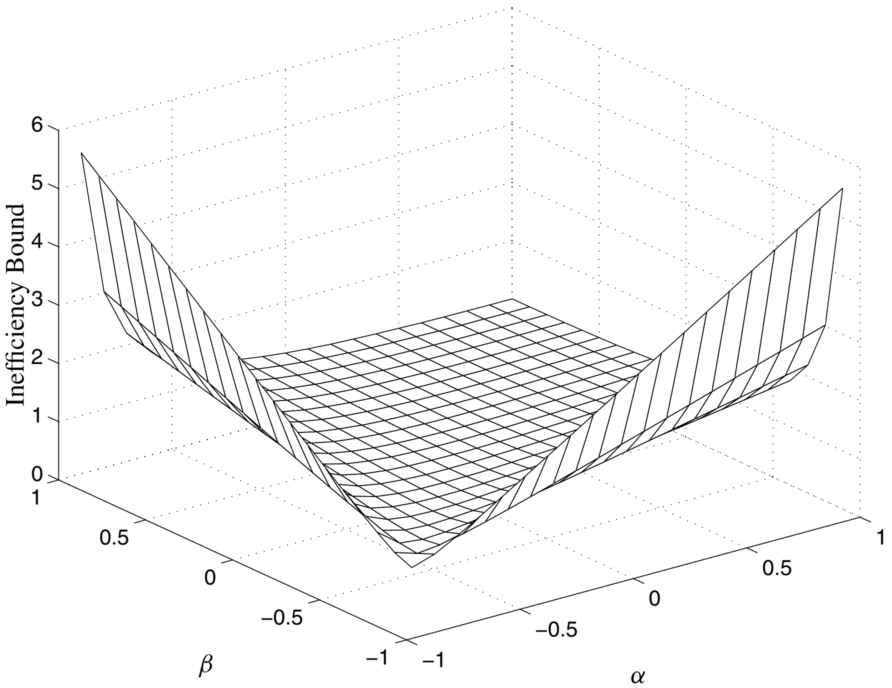

Figure 2.2: Measurement-Error Covariance Inefficiency Plot

case is 1.1681), as compared to when both are less than 1 (the average value for e in this case is 1.0175). This states that the estimate errors are worse when the assumed measurement-error covariance matrix is lower than the true covariance. This example clearly shows the influence of the measurement-error covariance on the performance characteristics of the estimates.

2.7.2 Ridge Estimation

As mentioned in §1.2.1 the inverse of $H^T H$ exists only if the number of linearly independent observations is equal to or greater than the number of unknowns, and if independent basis functions are used to form H. If the matrix $H^T H$ is close to being ill-conditioned, then the model is known as *weak multicollinear*. We can clearly see that weak multicollinearity may produce a large covariance in the estimated parameters. A strong multicollinearity exists if there are exact linear relations among the observations so that the rank of H equals n.[17, 18] This corresponds to the case of having linearly dependent rows in H. Another situation for $H^T H$ ill-conditioning is due to H having linearly independent columns, which occurs when the basis functions themselves are not independent of each other (e.g., choosing t, t^2 and $at+bt^2$, where a and b are constants, as basis functions leads to an ill-conditioned H matrix). Hoerl and Kennard[19] have proposed a class of estimators, called *ridge regression* estimators, that have a less total mean error than ordinary least squares (which is useful for the case of weak multicollinearity). However, as will be shown, the estimates are biased. Ridge estimation involves adding a positive constant, ϕ, to each diagonal element of $H^T H$, so that

$$\boxed{\hat{\mathbf{x}} = (H^T H + \phi I)^{-1} H^T \tilde{\mathbf{y}}} \tag{2.211}$$

Note the similarity between the ridge estimator and the Levenberg-Marquardt method in §1.6.3. Also note that even though the ridge estimator is a heuristic step motivated by numerical issues, comparing eqn. (2.79) to eqn. (2.211) leads to an equivalent relationship of formally treating $\hat{\mathbf{x}}_a = \mathbf{0}$ as an *a priori* estimate with associated covariance $Q = (1/\phi)I$. More generally, we may desire to use $\hat{\mathbf{x}}_a \neq \mathbf{0}$ and Q equal to some best estimate of the covariance of the errors in $\hat{\mathbf{x}}_a$.

We will first show that the ridge estimator produces biased estimates. Substituting eqn. (2.1) into eqn. (2.211) and taking the expectation leads to

$$E\left\{\hat{\mathbf{x}}\right\} = (H^T H + \phi I)^{-1} H^T H \mathbf{x} \tag{2.212}$$

Therefore, the bias is given by

$$\mathbf{b} \equiv E\left\{\hat{\mathbf{x}}\right\} - \mathbf{x} = \left[(H^T H + \phi I)^{-1} H^T H - I\right] \mathbf{x} \tag{2.213}$$

This can be simplified to yield

$$\mathbf{b} = -\phi (H^T H + \phi I)^{-1} \mathbf{x} \tag{2.214}$$

We clearly see that the ridge estimates are unbiased only when $\phi = 0$, which reduces to the standard least squares estimator.

Let us compute the covariance of the ridge estimator. Recall that the covariance is defined as

$$P \equiv E\left\{\hat{\mathbf{x}}\hat{\mathbf{x}}^T\right\} - E\left\{\hat{\mathbf{x}}\right\} E\left\{\hat{\mathbf{x}}\right\}^T \tag{2.215}$$

Assuming that $\mathbf{v}$ and $\mathbf{x}$ are uncorrelated leads to

$$P_{\text{ridge}} = (H^T H + \phi I)^{-1} H^T R H (H^T H + \phi I)^{-1} \tag{2.216}$$

Clearly, as ϕ increases the ridge covariance decreases, but at a price! The estimate becomes more biased, as seen in eqn. (2.214). We wish to find ϕ that minimizes the error $\hat{\mathbf{x}} - \mathbf{x}$, so that the estimate is as close to the truth as possible. A natural choice is to investigate the characteristics of the following matrix:

$$\Upsilon \equiv E\left\{(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^T\right\} \tag{2.217}$$

Note, this is *not* the covariance of the ridge estimate since $E\left\{\hat{\mathbf{x}}\right\} \neq \mathbf{x}$ in this case (therefore, the parallel axis theorem cannot be used). First, define

$$\Gamma \equiv (H^T H + \phi I)^{-1} \tag{2.218}$$

The following expectations can readily be derived

$$E\left\{\hat{\mathbf{x}}\hat{\mathbf{x}}^T\right\} = \Gamma\left(H^T R H + H^T H \mathbf{x}\mathbf{x}^T H^T H\right)\Gamma \tag{2.219}$$

$$E\left\{\mathbf{x}\hat{\mathbf{x}}^T\right\} = \mathbf{x}\mathbf{x}^T H^T H\,\Gamma \tag{2.220}$$

$$E\left\{\hat{\mathbf{x}}\mathbf{x}^T\right\} = \Gamma\, H^T H \mathbf{x}\mathbf{x}^T \tag{2.221}$$

Next, we make use of the following identities:

$$I - \Gamma H^T H = \phi\Gamma \tag{2.222}$$

and

$$\Gamma^{-1} - H^T H = \phi I \tag{2.223}$$

Hence, eqn. (2.217) becomes

$$\Upsilon = \Gamma\left(H^T R H + \phi^2 \mathbf{x}\mathbf{x}^T\right)\Gamma \tag{2.224}$$

We now wish to investigate the possibility of finding a range of ϕ that produces a lower Υ than the standard least squares covariance. In this analysis we will assume isotropic measurement errors so that $R = \sigma^2 I$. The least squares covariance can be manipulated using eqn. (2.218) to yield

$$\begin{aligned} P_{\text{ls}} &= \sigma^2 (H^T H)^{-1} \\ &= \sigma^2 \Gamma\left[\Gamma^{-1}(H^T H)^{-1}\Gamma^{-1}\right]\Gamma \\ &= \sigma^2 \Gamma\left[I + \phi(H^T H)^{-1}\right]\left[H^T H + \phi I\right]\Gamma \\ &= \sigma^2 \Gamma\left[\phi^2 (H^T H)^{-1} + 2\phi I + H^T H\right]\Gamma \end{aligned} \tag{2.225}$$

Using eqns. (2.216), (2.218), and (2.225), the condition for $P_{\text{ls}} - \Upsilon \geq 0$ is given by

$$\phi\Gamma\left\{\sigma^2\left[2I + \phi(H^T H)^{-1}\right] - \phi\mathbf{x}\mathbf{x}^T\right\}\Gamma \geq 0 \tag{2.226}$$

A sufficient condition for this inequality to hold true is $\phi \geq 0$ and

$$2\sigma^2 I - \phi\mathbf{x}\mathbf{x}^T \geq 0 \tag{2.227}$$

Left multiplying eqn. (2.227) by $\mathbf{x}^T$ and right multiplying the resulting expression by $\mathbf{x}$ leads to the following condition:

$$0 \leq \phi \leq \frac{2\sigma^2}{\mathbf{x}^T\mathbf{x}} \tag{2.228}$$

This guarantees that the inequality is satisfied; however, it is only a sufficient condition since we ignored the term $(H^T H)^{-1}$ in eqn. (2.226).

We can also choose to minimize the trace of Υ as well, which reduces the residual errors. Without loss in generality we can replace $H^T H$ with Λ, which is a diagonal matrix with elements given by the eigenvalues of $H^T H$. The trace of Υ is given by

$$\text{Tr}(\Upsilon) = \text{Tr}\left[(\Lambda + \phi I)^{-1}(\sigma^2 \Lambda + \phi^2 \mathbf{x}\mathbf{x}^T)(\Lambda + \phi I)^{-1}\right] \tag{2.229}$$

Therefore, we can now express the trace of Υ simply by

$$\text{Tr}(\Upsilon) = \sum_{i=1}^{n} \frac{\sigma^2 \lambda_i + \phi^2 x_i^2}{(\lambda_i + \phi)^2} \tag{2.230}$$

where λ_i is the i^{th} diagonal element of Λ. Minimizing eqn. (2.230) with respect to ϕ yields the following condition:

$$2\phi \sum_{i=1}^{n} \frac{\lambda_i x_i^2}{(\lambda_i + \phi)^3} - 2\sigma^2 \sum_{i=1}^{n} \frac{\lambda_i}{(\lambda_i + \phi)^3} = 0 \tag{2.231}$$

Since $\mathbf{x}$ is unknown, the optimal ϕ cannot be determined *a priori*.[20] One possible procedure to determine ϕ involves plotting each component of $\hat{\mathbf{x}}$ against ϕ, which is called a *ridge trace*. The estimates will stabilize at a certain value of ϕ. Also, the residual sum squares should be checked so that the condition in eqn. (2.228) is met.

Example 2.10: As an example of the performance tradeoffs in ridge estimation, we will consider a simple case with $x = 1.5$, $\sigma^2 = 2$, and $\lambda = 2$. A plot of the ridge variance, the least squares variance, the ridge residual sum squares, and the bias-squared quantities as a function of the ridge parameter ϕ is shown in Figure 2.3. From eqn. (2.231), using the given parameters, the optimal value for ϕ is 0.89. This is verified in Figure 2.3. From eqn. (2.228), the region where the residual sum squares is less than the least squares residual is given by $0 \le \phi \le 1.778$, which is again verified in Figure 2.3. As mentioned previously, this is a conservative condition (the actual upper bound is 3.200). From Figure 2.3, we also see that the ridge variance is always less than the least squares variance; however, the bias increases as ϕ increases.

Ridge estimation provides a powerful tool that can produce estimates that have smaller residual errors than traditional least squares. It is especially useful when $H^T H$ is close to being singular. However, in practical engineering applications involving dynamic systems biases are usually not tolerated, and thus the advantage of ridge estimation is diminished. In short, careful attention needs to be placed by the design engineer in order to weigh the possible advantages with the inevitable biased estimates in the analysis of the system. Alternatively, it may be possible to justify a

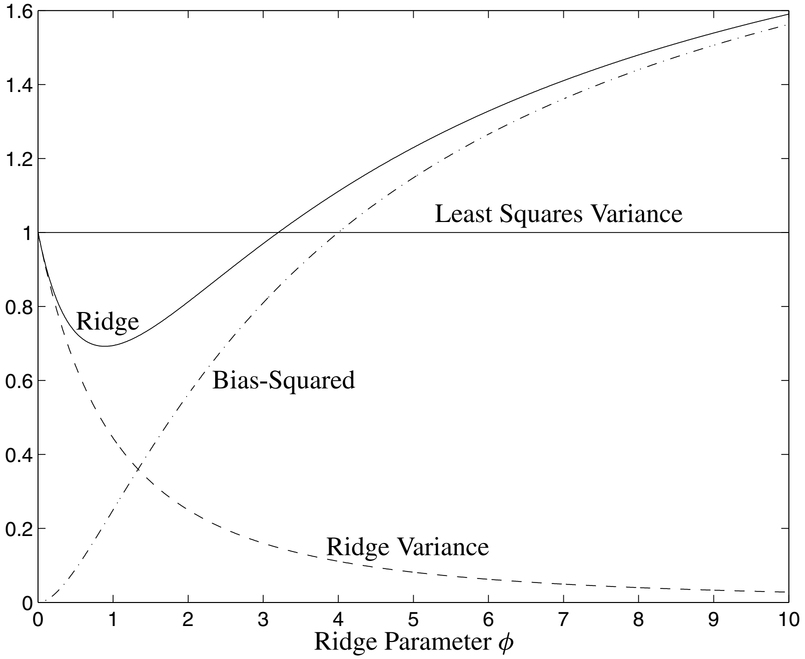

Figure 2.3: Ridge Estimation for a Scalar Case

particular ridge estimation process by using eqn. (2.79) for the case that a rigorous covariance Q is available for an *a priori* estimate $\hat{\mathbf{x}}_a$. Of course, in this theoretical setting, eqn. (2.79) is an unbiased estimator.

2.7.3 Total Least Squares

The standard least squares model in eqn. (2.1) assumes that there are no errors in the H matrix. Although this situation occurs in many systems, this assumption may not be always true. The least squares formulation in example 1.1 uses the measurements themselves in H, which contain random measurement errors. These "errors" were ignored in the least squares solution. Total least squares[21, 22] addresses errors in the H matrix, and can provide higher accuracy than ordinary least squares. In order to introduce this subject we begin by considering estimating a scalar parameter x:[22]

$$\tilde{\mathbf{y}} = \tilde{\mathbf{h}}\, x \tag{2.232}$$

with

$$\tilde{y}_i = y_i + v_i, \quad i = 1, 2, \ldots, m \tag{2.233a}$$

$$\tilde{h}_i = h_i + u_i, \quad i = 1, 2, \ldots, m \tag{2.233b}$$

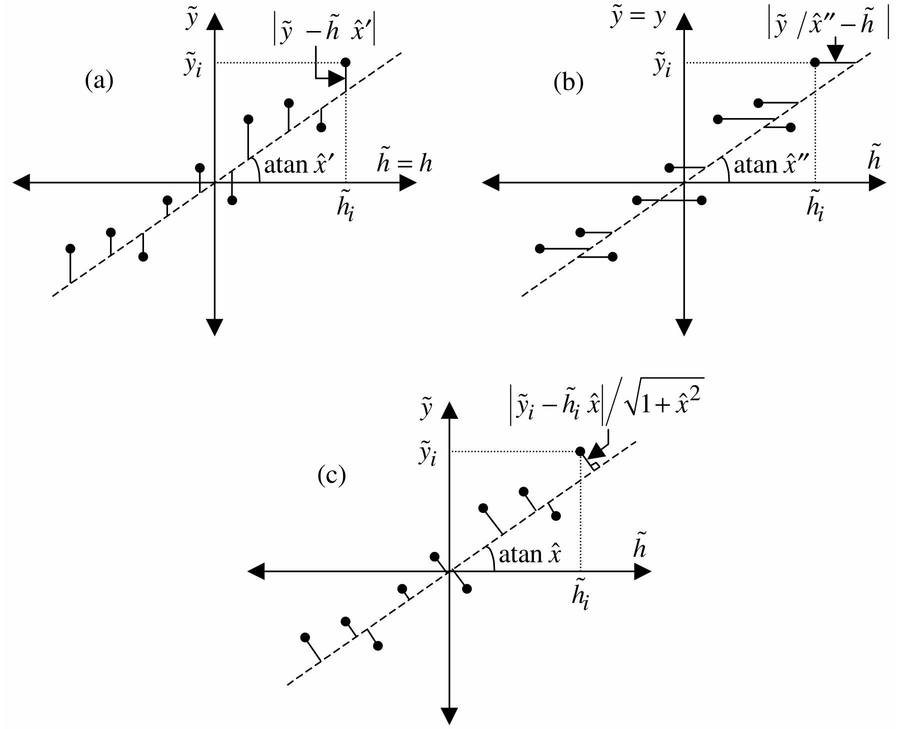

Figure 2.4: Geometric Interpretation of Total Least Squares

where v_i and u_i represent errors to the true values y_i and h_i, respectively.

When $u_i = 0$ then the estimate for x, denoted by $\hat{x}'$, is found by minimizing:

$$J(\hat{x}') = \sum_{i=1}^{m} (\tilde{y}_i - h_i\,\hat{x}')^2 \tag{2.234}$$

which yields

$$\hat{x}' = \left[\sum_{i=1}^{m} h_i^2\right]^{-1} \sum_{i=1}^{m} h_i\,\tilde{y}_i \tag{2.235}$$

The geometric interpretation of this result is shown by Case (a) in Figure 2.4. The residual is perpendicular to the $\tilde{h}$ axis. When $v_i = 0$ then the estimate for x, denoted by $\hat{x}''$, is found by the minimizing:

$$J(\hat{x}'') = \sum_{i=1}^{m} (y_i/\hat{x}'' - \tilde{h}_i)^2 \tag{2.236}$$

which yields

$$\hat{x}'' = \left[\sum_{i=1}^{m} \tilde{h}_i y_i\right]^{-1} \sum_{i=1}^{m} y_i^2 \tag{2.237}$$

The geometric interpretation of this result is shown by Case (b) in Figure 2.4. The residual is perpendicular to the $\tilde{y}$ axis. If the errors in both y_i and h_i have zero mean and have the same variance, then the total least squares estimate for x, denoted $\hat{x}$, is found by minimizing the sum of squared distances of the measurement points from the fitted line:

$$J(\hat{x}) = \sum_{i=1}^{m} (\tilde{y}_i - h_i \hat{x}')^2/(1+\hat{x}^2) \tag{2.238}$$

The solution for this minimization problem will be shown later. The geometric interpretation of this result is shown by Case (c) in Figure 2.4. The residual is now perpendicular to the fitted line. This geometric interpretation leads to the *orthogonal regression* approach in the total least squares problem.

For the general problem, the total least squares model is given by

$$\tilde{\mathbf{y}} = (H+U)\mathbf{x} + \mathbf{v} \tag{2.239}$$

where U represents the error to the model H. We assume $E\{\mathbf{v}\} = \mathbf{0}$ and $E\{\mathbf{v}\mathbf{v}^T\} = \sigma^2 I$ (i.e., the errors are isotropic). Furthermore we assume that the rows of U have zero mean with the same variance as the measurement error ($\sigma^2 I$). If this is not a valid assumption, then the matrix $\begin{bmatrix} H & \tilde{\mathbf{y}} \end{bmatrix}$ can be multiplied by an appropriate $m \times m$ matrix D such that the assumption is valid.[23]

The total least squares problem seeks an optimal estimate of $\mathbf{x}$ that minimizes

$$J = ||[\tilde{H}\ \tilde{\mathbf{y}}] - [\hat{H}\ \hat{\mathbf{y}}]||_F \tag{2.240}$$

where $\tilde{H} = H + U$ and $||\cdot||_F$ denotes the Frobenius norm (see §A.3). The solution for the total least squares problem is given by taking a singular value decomposition of the following augmented matrix:

$$\begin{bmatrix} \tilde{H} & \tilde{\mathbf{y}} \end{bmatrix} = USV^T \tag{2.241}$$

with $S = \text{diag}\begin{bmatrix} s_1 & \cdots & s_{n+1} \end{bmatrix}$. The total least squares solution is then given by[24]

$$\boxed{\hat{\mathbf{x}}_{\text{TLS}} = (\tilde{H}^T \tilde{H} - s_{n+1}^2 I)^{-1} \tilde{H}^T \tilde{\mathbf{y}}} \tag{2.242}$$

Notice the resemblance to ridge estimation in §2.7.2, but here the positive multiple is subtracted from $\tilde{H}^T \tilde{H}$. Therefore, the total least squares problem is a *deregularization* of the least squares problem, which means that it is always worse conditioned than the ordinary least squares problem.

Total least squares has been shown to provide parameter error accuracy gains of 10 to 15 percent in typical applications.[24] In order to quantify the bounds on the difference between total least squares and ordinary least squares we begin by using the following identity:

$$(\tilde{H}^T\tilde{H} - s_{n+1}^2 I)\hat{\mathbf{x}}_{\text{LS}} = \tilde{H}^T\tilde{\mathbf{y}} - s_{n+1}^2\hat{\mathbf{x}}_{\text{LS}} \tag{2.243}$$

Subtracting eqn. (2.243) from eqn. (2.242) leads to

$$\hat{\mathbf{x}}_{\text{TLS}} - \hat{\mathbf{x}}_{\text{LS}} = s_{n+1}^2(\tilde{H}^T\tilde{H} - s_{n+1}^2 I)^{-1}\hat{\mathbf{x}}_{\text{LS}} \tag{2.244}$$

Using the norm inequality now leads to:

$$\frac{||\hat{\mathbf{x}}_{\text{TLS}} - \hat{\mathbf{x}}_{\text{LS}}||}{||\hat{\mathbf{x}}_{\text{LS}}||} \le \frac{s_{n+1}^2}{\bar{s}_n^2 - s_{n+1}^2} \tag{2.245}$$

where $\bar{s}_n$ is the smallest singular value of $\tilde{H}$ and the assumption $\bar{s}_n > s_{n+1}$ must be valid. The accuracy of total least squares will be more pronounced when the ratio of the singular values $\bar{s}_n$ and s_{n+1} is large. The "errors-in-variables" estimator shown in Ref. [25] coincides with the total least squares solution. This indicates that the total least squares estimate is a strongly consistent estimate for large samples, which leads to an asymptotic unbiasedness property. Ordinary least squares with errors in H produces biased estimates as the sample size increases. However, the covariance of total least squares is larger than the ordinary least squares covariance, but by increasing the noise in the measurements the bias of ordinary least squares becomes more important and even the dominating term.[22] Several aspects and properties of the total least squares problem can be found in the references cited in this section.

Example 2.11: We will show the advantages of total least squares by re-considering the problem of estimating the parameters of a simple dynamic system shown in example 1.1. To compare the accuracy of total least squares with ordinary least squares we will use the square root of the diagonal elements of mean-squared-error (MSE) matrix, defined as

$$\begin{aligned}
\text{MSE} &= E\left\{\left(\hat{\mathbf{x}} - \mathbf{x}\right)\left(\hat{\mathbf{x}} - \mathbf{x}\right)^T\right\} \\
&= E\left\{\left(\hat{\mathbf{x}} - E\{\hat{\mathbf{x}}\}\right)\left(\hat{\mathbf{x}} - E\{\hat{\mathbf{x}}\}\right)^T\right\} + \left\{\left(E\{\hat{\mathbf{x}}\} - \mathbf{x}\right)\left(E\{\hat{\mathbf{x}}\} - \mathbf{x}\right)^T\right\} \\
&= \text{cov}\{\hat{\mathbf{x}}\} + \text{squared bias}\{\hat{\mathbf{x}}\}
\end{aligned}$$

For this particular problem it is known that u is given by an impulse input with magnitude $10/\Delta t$ (i.e., $u_1 = 10/\Delta t$ and $u_k = 0$ for $k \ge 2$). A total of 10 seconds is considered with sampling intervals ranging from $\Delta t = 2$ seconds down to $\Delta t = 0.001$ seconds. Synthetic measurements are again generated with $\sigma = 0.08$. This example tests the accuracy of both approaches for various measurement sample lengths (i.e., from 5 samples when $\Delta t = 2$ to 10,000 samples when $\Delta t = 0.001$). For each simulation 1,000 runs were performed each with different random number seeds. Results for $\hat{\Phi}$ are given in the following table:

Δt	bias$\{\hat{\Phi}\}_{LS}$	bias$\{\hat{\Phi}\}_{TLS}$	$\sqrt{\text{MSE}\{\hat{\Phi}\}_{LS}}$	$\sqrt{\text{MSE}\{\hat{\Phi}\}_{TLS}}$
2	3.12×10^{-4}	3.89×10^{-4}	1.82×10^{-2}	1.83×10^{-2}
1	5.52×10^{-4}	2.43×10^{-4}	1.12×10^{-2}	1.12×10^{-2}
0.5	1.03×10^{-3}	3.67×10^{-4}	6.36×10^{-3}	6.28×10^{-3}
0.1	1.24×10^{-3}	9.68×10^{-5}	1.99×10^{-3}	1.54×10^{-3}
0.05	1.23×10^{-3}	2.30×10^{-5}	1.47×10^{-3}	7.90×10^{-4}
0.01	1.26×10^{-3}	7.08×10^{-6}	1.28×10^{-3}	1.62×10^{-4}
0.005	1.27×10^{-3}	3.48×10^{-6}	1.27×10^{-3}	8.26×10^{-5}
0.001	1.28×10^{-3}	5.32×10^{-7}	1.27×10^{-3}	1.60×10^{-5}

Results for $\hat{\Gamma}$ are given in the following table:

Δt	bias$\{\hat{\Gamma}\}_{LS}$	bias$\{\hat{\Gamma}\}_{TLS}$	$\sqrt{\text{MSE}\{\hat{\Gamma}\}_{LS}}$	$\sqrt{\text{MSE}\{\hat{\Gamma}\}_{TLS}}$
2	1.37×10^{-4}	1.11×10^{-4}	8.37×10^{-3}	8.78×10^{-3}
1	1.32×10^{-4}	6.24×10^{-5}	6.64×10^{-3}	6.71×10^{-3}
0.5	1.29×10^{-4}	2.25×10^{-5}	4.76×10^{-3}	4.76×10^{-3}
0.1	1.52×10^{-5}	2.11×10^{-5}	1.07×10^{-3}	1.07×10^{-3}
0.05	2.71×10^{-5}	2.87×10^{-5}	5.61×10^{-4}	5.62×10^{-4}
0.01	7.04×10^{-6}	7.10×10^{-6}	1.12×10^{-4}	1.13×10^{-4}
0.05	2.02×10^{-6}	2.00×10^{-6}	5.90×10^{-5}	5.91×10^{-5}
0.001	1.79×10^{-7}	2.78×10^{-7}	1.10×10^{-5}	1.11×10^{-5}

These tables indicate that when using a small sample size ordinary least squares and total least squares have the same accuracy. However, as the sampling interval decreases (i.e., giving more measurements) the bias in $\hat{\Phi}$ increases using ordinary least squares, but substantially decreases using total least squares. Also, the bias is the dominating term in the MSE when the sample size is large. Results for $\hat{\Gamma}$ indicate that the ordinary least squares estimate is comparable to the total least squares estimate. This is due to the fact that u contains no errors. Nevertheless, this example clearly shows that improvements can be made using total least squares.

2.8 Summary

In this chapter we have presented several approaches to establish a class of linear estimation algorithms, and we have developed certain important properties of the weighting matrix used in weighted least squares. The end products of the developments for minimum variance estimation in §2.1.1 and maximum likelihood estimation in §2.3 are seen to be equivalent for Gaussian measurement errors to the linear weighted least squares results of §1.2.2, with interpretation of the weight matrix as the measurement-error covariance matrix. An interesting result is that several different theoretical/conceptual estimation approaches give the same estimator. In particular, when weighing the advantages and disadvantages of each approach one realizes that maximum likelihood provides a solution more directly than minimum variance, since a constrained optimization problem is not required. Therefore, in practice, maximum likelihood estimation is usually preferred over minimum variance. Several useful properties were also derived in this chapter, including unbiased estimates and the Cramér-Rao inequality. In estimation of dynamic systems, an unbiased estimate is always preferred, if obtainable, over a biased estimate. Also, an efficient estimator, which is achieved if the equality in the Cramér-Rao inequality is satisfied, gives the lowest estimation error possible from a statistical point of view. This allows the design engineer to quantify the performance of an estimation algorithm using a covariance analysis on the expected performance.

The interpretation of the *a priori* estimates in §2.1.2 is given as a measurement subset in the sequential least squares developments of §1.3. Several other approaches, such as maximum *a posteriori* estimation and minimum risk estimation of §2.6 were shown to be equivalent to the minimum variance solution of §2.1.2. Each of these approaches provides certain illuminations and useful insights. Maximum *a posteriori* estimation is usually preferred over the other approaches since it follows many of the same principles and properties of maximum likelihood estimation, and in fact reduces to the maximum likelihood estimate if the *a priori* distribution is uniform or for large samples. The Cramér-Rao bound for *a priori* estimation was also shown, which again provides a lower bound on the estimation error.

In §2.5 a discussion on the nonuniqueness of the weight matrix was given. It should be noted that specification and calculations involving the weight matrices are the source of most practical difficulties encountered in applications. Additionally, an analysis of errors in the assumed measurement-error covariance matrix was shown in §2.7.1. This analysis can be useful to quantify the expected performance of the estimate in the face of an incorrectly defined measurement-error covariance matrix. Ridge estimation, shown in §2.7.2, is useful for the case of weak multicollinear systems. This case involves the near ill-conditioning of the matrix to be inverted in the least squares solutions. It has also been established that the ridge estimate covariance is less than the least squares estimate covariance. However, if the least squares solution is well posed, then the advantage of a lower covariance is strongly outweighed by the inevitable biased estimate in ridge estimation. Also, a connection between ridge estimation and *a priori* state estimation has been established by noting that resemblance of the ridge parameter to the *a priori* covariance. Finally, total least squares, shown in §2.7.3, can give significant improvements in the accuracy of the estimates over ordinary least squares if errors are present in the model matrix. This approach synthesizes an optimal methodology for solving a variety of problems in many dynamic system applications.

A summary of the key formulas presented in this chapter is given below.

- Gauss-Markov Theorem

$$\tilde{\mathbf{y}} = H\mathbf{x} + \mathbf{v}$$
$$E\{\mathbf{v}\} = \mathbf{0}, \quad E\left\{\mathbf{v}\,\mathbf{v}^T\right\} = R$$
$$\hat{\mathbf{x}} = (H^T R^{-1} H)^{-1} H^T R^{-1} \tilde{\mathbf{y}}$$

- *A priori* Estimation

$$\tilde{\mathbf{y}} = H\mathbf{x} + \mathbf{v}$$
$$E\{\mathbf{v}\} = \mathbf{0}, \quad E\left\{\mathbf{v}\,\mathbf{v}^T\right\} = R$$
$$\hat{\mathbf{x}}_a = \mathbf{x} + \mathbf{w}$$
$$E\{\mathbf{w}\} = \mathbf{0}, \quad E\left\{\mathbf{w}\,\mathbf{w}^T\right\} = Q$$
$$\hat{\mathbf{x}} = \left(H^T R^{-1} H + Q^{-1}\right)^{-1} \left(H^T R^{-1} \tilde{\mathbf{y}} + Q^{-1} \hat{\mathbf{x}}_a\right)$$

- Unbiased Estimates

$$E\left\{\hat{\mathbf{x}}_k(\tilde{\mathbf{y}})\right\} = \mathbf{x} \quad \text{for all } k$$

- Maximum Likelihood Estimation

$$L(\tilde{\mathbf{y}};\mathbf{x}) = \prod_{i=1}^{p} f_i(\tilde{\mathbf{y}};\mathbf{x})$$
$$\left\{\frac{\partial}{\partial \mathbf{x}} \ln\left[L(\tilde{\mathbf{y}};\mathbf{x})\right]\right\}\Bigg|_{\hat{\mathbf{x}}} = \mathbf{0}$$

- Cramér-Rao Inequality

$$P \equiv E\left\{\left(\hat{\mathbf{x}} - \mathbf{x}\right)\left(\hat{\mathbf{x}} - \mathbf{x}\right)^T\right\} \geq F^{-1}$$
$$F = -E\left\{\frac{\partial^2}{\partial \mathbf{x}\,\partial \mathbf{x}^T} \ln f(\tilde{\mathbf{y}};\mathbf{x})\right\}$$

- Bayes Rule

$$f(\mathbf{x}|\tilde{\mathbf{y}}) = \frac{f(\tilde{\mathbf{y}}|\mathbf{x})\,f(\mathbf{x})}{f(\tilde{\mathbf{y}})}$$

- Maximum *A Posteriori* Estimation

$$J_{\text{MAP}}(\hat{\mathbf{x}}) = \ln\left[L(\tilde{\mathbf{y}}|\hat{\mathbf{x}})\right] + \ln\left[f(\hat{\mathbf{x}})\right]$$

- Cramér-Rao Inequality for Bayesian Estimators

$$\begin{aligned} P &\equiv E\left\{\left(\hat{\mathbf{x}}-\mathbf{x}\right)\left(\hat{\mathbf{x}}-\mathbf{x}\right)^T\right\} \\ &\geq \left[F + E\left\{\left[\frac{\partial}{\partial \mathbf{x}}\ln f(\mathbf{x})\right]\left[\frac{\partial}{\partial \mathbf{x}}\ln f(\mathbf{x})\right]^T\right\}\right]^{-1} \end{aligned}$$

- Minimum Risk Estimation

$$J_{\text{MR}}(\mathbf{x}^*) = \int_{-\infty}^{\infty} c(\mathbf{x}^*|\mathbf{x})\frac{f(\tilde{\mathbf{y}}|\mathbf{x})\,f(\mathbf{x})}{f(\tilde{\mathbf{y}})}\,d\mathbf{x}$$

$$c(\mathbf{x}|\mathbf{x}^*) = \frac{1}{2}(\mathbf{x}^* - \mathbf{x})^T S(\mathbf{x}^* - \mathbf{x})$$

$$\hat{\mathbf{x}} = \int_{-\infty}^{\infty} \mathbf{x}\frac{f(\tilde{\mathbf{y}}|\mathbf{x})\,f(\mathbf{x})}{f(\tilde{\mathbf{y}})}\,d\mathbf{x}$$

- Inefficiency for Covariance Errors

$$e = \frac{\det\left[(H^T\tilde{R}^{-1}H)^{-1}H^T\tilde{R}^{-1}R\,\tilde{R}^{-1}H(H^T\tilde{R}^{-1}H)^{-1}\right]}{\det\left[(H^TR^{-1}H)^{-1}\right]}$$

- Ridge Estimation

$$\hat{\mathbf{x}} = (H^TH + \phi I)^{-1}H^T\tilde{\mathbf{y}}$$

- Total Least Squares

$$\begin{gathered} \tilde{\mathbf{y}} = \tilde{H}\mathbf{x} + \mathbf{v} \\ \left[\tilde{H}\ \tilde{\mathbf{y}}\right] = USV^T \\ S = \text{diag}\left[s_1\ \cdots\ s_{n+1}\right] \\ \hat{\mathbf{x}}_{\text{TLS}} = (\tilde{H}^T\tilde{H} - s_{n+1}^2 I)^{-1}\tilde{H}^T\tilde{\mathbf{y}} \end{gathered}$$

Exercises

2.1 Consider estimating a constant unknown variable x, which is measured twice with some error

$$\begin{aligned} \tilde{y}_1 &= x + v_1 \\ \tilde{y}_2 &= x + v_2 \end{aligned}$$

where the random errors have the following properties:

$$E\{v_1\} = E\{v_2\} = E\{v_1 v_2\} = 0$$
$$E\left\{v_1^2\right\} = 1$$
$$E\left\{v_2^2\right\} = 4$$

Perform a weighted least squares solution with $H = \begin{bmatrix}1 & 1\end{bmatrix}^T$ for the following two cases:

$$W = \frac{1}{2}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

and

$$W = \frac{1}{4}\begin{bmatrix}4 & 0\\ 0 & 1\end{bmatrix}$$

Compute the variance of the estimation error (i.e. $E\left\{(x-\hat{x})^2\right\}$) and compare the results.

2.2 Write a simple computer program to simulate measurements of some discretely measured process

$$\tilde{y}_j = x_1 + x_2 \sin(10t_j) + x_3 e^{2t_j^2} + v_j, \quad j = 1, 2, \ldots, 11$$

with t_j sampled every 0.1 seconds. The true values (x_1, x_2, x_3) are $(1, 1, 1)$ and the measurement errors are synthetic Gaussian random variables with zero mean. The measurement-error covariance matrix is diagonal with

$$R = E\left\{\mathbf{v}\mathbf{v}^T\right\} = \text{diag}\left[\sigma_1^2 \; \sigma_2^2 \; \cdots \; \sigma_{11}^2\right]$$

where

$$\begin{array}{llll} \sigma_1 = 0.001 & \sigma_2 = 0.002 & \sigma_3 = 0.005 & \sigma_4 = 0.010 \\ \sigma_5 = 0.008 & \sigma_6 = 0.002 & \sigma_7 = 0.010 & \sigma_8 = 0.007 \\ \sigma_9 = 0.020 & \sigma_{10} = 0.006 & \sigma_{11} = 0.001 & \end{array}$$

You are also given the *a priori* $\mathbf{x}$-estimates

$$\hat{\mathbf{x}}_a^T = (1.01, 0.98, 0.99)$$

and associated *a priori* covariance matrix

$$Q = \begin{bmatrix} 0.001 & 0 & 0 \\ 0 & 0.001 & 0 \\ 0 & 0 & 0.001 \end{bmatrix}$$

Your tasks are as follows:

(A) Use the minimal variance estimation version of the normal equations

$$\hat{\mathbf{x}} = P\left(H^T R^{-1} \tilde{\mathbf{y}} + Q^{-1} \hat{\mathbf{x}}_a\right)$$

to compute the parameter estimates and estimate covariance matrix

$$P = \left(H^T R^{-1} H + Q^{-1}\right)^{-1}$$

with the j^{th} row of H given by $\begin{bmatrix} 1 & \sin(10t_j) & e^{2t_j^2} \end{bmatrix}$. Calculate the mean and standard deviation of the residual

$$r_j = \tilde{y}_j - \left(\hat{x}_1 + \hat{x}_2 \sin(10t_j) + \hat{x}_3 e^{2t_j^2}\right)$$

as

$$r = \frac{1}{11} \sum_{j=1}^{11} r_j$$

$$\sigma_r = \left[\frac{1}{10} \sum_{j=1}^{11} r_j^2\right]^{\frac{1}{2}}$$

(B) Do a parametric study in which you hold the *a priori* estimate covariance Q fixed, but vary the measurement-error covariance according to

$$R' = \alpha R$$

with $\alpha = 10^{-3}$, 10^{-2}, 10^{-1}, 10, 10^2, 10^3. Study the behavior of the calculated results for the estimates $\hat{\mathbf{x}}$, the estimate covariance matrix P, and mean r and standard deviation σ_r of the residual.

(C) Do a parametric study in which R is held fixed, but Q is varied according to

$$Q' = \alpha Q$$

with α taking the same values as in (B). Compare the results for the estimates $\hat{\mathbf{x}}$, the estimate covariance matrix P, and mean r and standard deviation σ_r of the residual with those of part (B).

2.3 Suppose that $\mathbf{v}$ in exercise 1.3 is a constant vector (i.e., a *bias error*). Evaluate the loss function (2.111) in terms of v_i only and discuss how the value of the loss function changes with a bias error in the measurements instead of a zero mean assumption.

2.4 Consider the following constrained minimization problem:

$$\begin{aligned} &\text{minimize} \quad J(\hat{\mathbf{x}}) = \left(\tilde{\mathbf{y}} - H\hat{\mathbf{x}}\right)^T R^{-1} \left(\tilde{\mathbf{y}} - H\hat{\mathbf{x}}\right) \\ &\text{subject to} \quad S\hat{\mathbf{x}} = \mathbf{z} \end{aligned}$$

where $\mathbf{z}$ and S are known (the matrix S has dimensions less than or equal to the dimension of the vector $\hat{\mathbf{x}}$). You are to perform the following tasks:

(A) Determine the least squares estimate $\hat{\mathbf{x}}$.

(B) Check to see if the estimator is unbiased.

(C) Derive the estimate covariance matrix.

(D) Determine the estimate when S is square with rank n.

2.5 A "Monte Carlo" approach to calculating covariance matrices is often necessary for nonlinear problems. The algorithm has the following structure: Given a functional dependence of two sets of random variables in the form

$$z_i = F_i(y_1, y_2, \ldots, y_m), \quad i = 1, 2, \ldots, n$$

where the y_j are random variables whose joint probability density function is known and the F_i are generally nonlinear functions. The Monte Carlo approach requires that the probability density function of y_j be sampled many times to calculate corresponding samples of the z_i joint distribution. Thus if the k^{th} particular sample ("simulated measurement") of the y_j values is denoted as

$$(\tilde{y}_{1k}, \tilde{y}_{2k}, \ldots, \tilde{y}_{mk}), \quad k = 1, 2, \ldots, M$$

then the corresponding z_i sample is calculated as

$$z_{ik} = F_i(\tilde{y}_{1k}, \tilde{y}_{2k}, \ldots, \tilde{y}_{mk}), \quad k = 1, 2, \ldots, M$$

The first two moments of z_i's joint density function are then approximated by

$$\mu_i = E\{z_{ik}\} \simeq \frac{1}{M}\sum_{k=1}^{M} z_{ik}$$

and

$$\hat{R} = E\left\{(\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^T\right\} \simeq \frac{1}{M-1}\sum_{k=1}^{M}[\mathbf{z}_k - \boldsymbol{\mu}][\mathbf{z}_k - \boldsymbol{\mu}]^T$$

where

$$\mathbf{z}_k^T \equiv (z_{1k}, z_{2k}, \ldots, z_{nk})$$
$$\boldsymbol{\mu}^T \equiv (\mu_1, \mu_2, \ldots, \mu_n)$$

The Monte Carlo approach can be used to experimentally verify the interpretation of $P = (H^T R^{-1} H)^{-1}$ as the $\hat{\mathbf{x}}$ covariance matrix in the minimal variance estimate

$$\hat{\mathbf{x}} = P H^T R^{-1} \tilde{\mathbf{y}}$$

To carry out this experiment, use the model in exercise 2.2 to simulate $M = 100$ sets of y-measurements. For each set (e.g., the k^{th}) of the measurements, the corresponding $\hat{\mathbf{x}}$ follows as

$$\hat{\mathbf{x}}_k = P H^T R^{-1} \tilde{\mathbf{y}}_k$$

Then the $\hat{\mathbf{x}}$ mean and covariance matrices can be approximated by

$$\boldsymbol{\mu}_x = E\left\{\hat{\mathbf{x}}\right\} \simeq \frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{x}}_k$$

and

$$\hat{R}_{xx} = E\left\{(\hat{\mathbf{x}} - \boldsymbol{\mu}_x)(\hat{\mathbf{x}} - \boldsymbol{\mu}_x)^T\right\} \simeq \frac{1}{M-1}\sum_{k=1}^{M}[\hat{\mathbf{x}}_k - \boldsymbol{\mu}_x][\hat{\mathbf{x}}_k - \boldsymbol{\mu}_x]^T$$

In your simulation $\hat{R}_{xx}$ should be compared element-by-element with the covariance $P = (H^T R^{-1} H)^{-1}$, whereas $\boldsymbol{\mu}_x$ should compare favorably with the true values $\mathbf{x}^T = (1, 1, 1)$.

2.6 Let $\tilde{\mathbf{y}} \sim \mathcal{N}(\boldsymbol{\mu}, R)$. Show that

$$\hat{\boldsymbol{\mu}} = \frac{1}{m} \sum_{k=1}^{M} \tilde{\mathbf{y}}_i$$

is an efficient estimator for the mean.

2.7 Consider estimating a constant unknown variable x, which is measured twice with some error

$$\begin{aligned} \tilde{y}_1 &= x + v_1 \\ \tilde{y}_2 &= x + v_2 \end{aligned}$$

where the random errors have the following properties:

$$\begin{aligned} E\{v_1\} &= E\{v_2\} = 0 \\ E\left\{v_1^2\right\} &= \sigma_1^2 \\ E\left\{v_2^2\right\} &= \sigma_2^2 \end{aligned}$$

The errors follow a bivariate normal distribution with joint density function given by

$$f(v_1, v_2) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{v_1^2}{\sigma_1^2} - \frac{2\rho v_1 v_2}{\sigma_1\sigma_2} + \frac{v_2^2}{\sigma_2^2}\right)\right]$$

where the correlation coefficient, ρ, is defined as

$$\rho \equiv \frac{E\{v_1 v_2\}}{\sigma_1\sigma_2}$$

Derive the maximum likelihood estimate for x. Also, how does the estimate change when $\rho = 0$?

2.8 Suppose that z_1 is the mean of a random sample of size m from a normal distributed system with mean μ and variance σ_1^2, and z_2 is the mean of a random sample of size m from a normal distributed system with mean μ and variance σ_2^2. Show that $\hat{\mu} = \alpha z_1 + (1-\alpha) z_2$, where $0 \le \alpha \le 1$, is an unbiased estimate of μ. Also, show that the variance of the estimate is minimum when $\alpha = \sigma_2^2(\sigma_1^2 + \sigma_2^2)^{-1}$.

2.9 Show that if $\hat{x}$ is an unbiased estimate of x and $\text{var}\{\hat{x}\}$ does not equal 0, then $\hat{x}^2$ is not an unbiased estimate of x^2.

2.10 If $\hat{x}$ is an estimate of x, its bias is $b = E\{\hat{x}\} - x$. Show that $E\left\{(\hat{x} - x)^2\right\} = \text{var}\{\hat{x}\} + b^2$.

2.11 Prove that the *a priori* estimator given in eqn. (2.47) is unbiased when $MH + N = I$ and $\mathbf{n} = \mathbf{0}$.

2.12 Prove that the Cramér-Rao inequality given by eqn. (2.112) achieves the equality if and only if

$$\left[\frac{\partial}{\partial \mathbf{x}} \ln f(\tilde{\mathbf{y}}; \mathbf{x})\right] = c\,(\mathbf{x} - \hat{\mathbf{x}})$$

where c is independent of $\mathbf{x}$ and $\tilde{\mathbf{y}}$.

2.13 Suppose that an estimator of a non-random scalar x is biased, with bias denoted by $b(x)$. Show that a lower bound on the variance of the estimate $\hat{x}$ is given by

$$\text{var}(\hat{x} - x) \geq \left(1 - \frac{db}{dx}\right)^2 J^{-1}$$

where

$$J = E\left\{\left[\frac{\partial}{\partial x} \ln f(\tilde{\mathbf{y}}; x)\right]^2\right\}$$

and

$$b(x) \equiv \int_{-\infty}^{\infty} (x - \hat{x})\, f(\tilde{\mathbf{y}}; x)\, d\tilde{\mathbf{y}}$$

2.14 Prove that the estimate for the covariance in example 2.4 is biased. Also, what is the unbiased estimate?

2.15 Prove that eqn. (2.114) is equivalent to eqn. (2.113).

2.16 Perform a simulation of the parameter identification problem shown in example 2.6 with $B = 10$ and varying σ for the measurement noise. Compare the nonlinear least squares solution to the linear approach for various noise levels. Also, check the performance of the two approaches by comparing P with $\mathcal{P}$. At what measurement noise level does the linear solution begin to degrade from the nonlinear least squares solution?

2.17 ♣ In example 2.6 an expression for the variance of the new measurement noise, denoted by ϵ_k, is derived. Prove the following expression:

$$E\left\{\left(\frac{v_k}{B\,e^{a\,t_k}} - \frac{v_k^2}{2\,B^2 e^{2\,a\,t_k}}\right)^2\right\} = \frac{\sigma^2}{B^2 e^{2\,a\,t_k}} + \frac{3\,\sigma^4}{4\,B^4 e^{4\,a\,t_k}}$$

Hint: use the theory behind χ^2 distributions.

2.18 ♣ Prove the inequality in eqn. (2.159).

2.19 The parallel axis theorem was used several times in this chapter to derive the covariance expression, e.g., in eqn. (2.181). Prove the following identity:

$$E\left\{(\mathbf{x} - E\{\mathbf{x}\})\,(\mathbf{x} - E\{\mathbf{x}\})^T\right\} = E\left\{\mathbf{x}\,\mathbf{x}^T\right\} - E\{\mathbf{x}\}\,E\{\mathbf{x}\}^T$$

2.20 Fully derive the density function given in eqn. (2.183).

2.21 Show that $\mathbf{e}^T R^{-1}\mathbf{e}$ is equivalent to $\mathrm{Tr}\left(R^{-1}E\right)$ with $E = \mathbf{e}\,\mathbf{e}^T$.

2.22 Prove that $E\left\{\mathbf{x}^T A\mathbf{x}\right\} = \boldsymbol{\mu}^T A\boldsymbol{\mu} + \mathrm{Tr}\,(A\,\Xi)$, where $E\{\mathbf{x}\} = \boldsymbol{\mu}$ and $\mathrm{cov}(\mathbf{x}) = \Xi$.

2.23 Prove the following results for the *a priori* estimator in eqn. (2.186):

$$E\left\{\mathbf{x}\,\hat{\mathbf{x}}^T\right\} = E\left\{\hat{\mathbf{x}}\,\hat{\mathbf{x}}^T\right\}$$

$$\left(H^T R^{-1} H + Q^{-1}\right)^{-1} = E\left\{\mathbf{x}\,\mathbf{x}^T\right\} - E\left\{\hat{\mathbf{x}}\,\hat{\mathbf{x}}^T\right\}$$

$$E\left\{\mathbf{x}\,\mathbf{x}^T\right\} \geq E\left\{\hat{\mathbf{x}}\,\hat{\mathbf{x}}^T\right\}$$

2.24 Consider the 2×2 case for $\tilde{R}$ and R in eqn. (2.199). Verify that the inefficiency e in eqn. (2.210) is bounded by

$$1 \leq e \leq \frac{(\lambda_{\max} + \lambda_{\min})^2}{4\lambda_{\max}\lambda_{\min}}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of the matrix $\tilde{R}^{-1/2} R\,\tilde{R}^{-1/2}$. Note, this inequality does not generalize to the case where $m \geq 3$.

2.25 ♣ An alternative to minimizing the trace of Υ in §2.7.2 is to minimize the generalized cross-validation (GRV) error prediction,[26] given by

$$\hat{\sigma}^2 = \frac{m\,\tilde{\mathbf{y}}^T \mathcal{P}^2 \tilde{\mathbf{y}}}{\mathrm{Tr}(P)^2}$$

where m is the dimension of the vector $\tilde{y}$ and $\mathcal{P}$ is a projection matrix, given by

$$\mathcal{P} = I - H(H^T H + \phi I)^{-1} H^T$$

Determine the minimum of the GRV error, as a function of the ridge parameter ϕ. Also, prove that $\mathcal{P}$ is a projection matrix.

2.26 Consider the following model:

$$y = x_1 + x_2 t + x_3 t^2$$

Create a set of 101 noise-free observations at 0.01-second intervals with $x_1 = 3$, $x_2 = 2$, and $x_3 = 1$. Form the H matrix to be used in least squares with basis functions given by $\left\{1,\, t,\, t^2,\, 2t + 3t^2\right\}$. Show that H is rank deficient. Use the ridge estimator in eqn. (2.211) to determine the parameter estimates with the aforementioned basis functions. How does varying ϕ affect the solution?

2.27 Write a computer program to reproduce the total least squares results shown in example 2.11.

2.28 ♣ Let the last column (denoted by $\mathbf{v}$) of the matrix of V in eqn. (2.241) be partitioned into the $n \times 1$ vector ($\mathbf{g}$) and the remaining scalar component (γ), such that $\mathbf{v} = \left[\mathbf{g}^T \ \gamma\right]^T$. Show that the total least squares estimate in eqn. (2.242) is equivalent to $\hat{\mathbf{x}}_{\text{TLS}} = -\gamma^{-1}\mathbf{g}$.

References

[1] Berry, D.A. and Lingren, B.W., *Statistics, Theory and Methods*, Brooks/Cole Publishing Company, Pacific Grove, CA, 1990.

[2] Goldstein, H., *Classical Mechanics*, Addison-Wesley Publishing Company, Reading, MA, 2nd ed., 1980.

[3] Baruh, H., *Analytical Dynamics*, McGraw-Hill, Boston, MA, 1999.

[4] Devore, J.L., *Probability and Statistics for Engineering and Sciences*, Duxbury Press, Pacific Grove, CA, 1995.

[5] Sorenson, H.W., *Parameter Estimation, Principles and Problems*, Marcel Dekker, New York, NY, 1980.

[6] Sage, A.P. and Melsa, J.L., *Estimation Theory with Applications to Communications and Control*, McGraw-Hill Book Company, New York, NY, 1971.

[7] Freund, J.E. and Walpole, R.E., *Mathematical Statistics*, Prentice Hall, Englewood Cliffs, NJ, 4th ed., 1987.

[8] Cramér, H., *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1946.

[9] Fisher, R.A., *Contributions to Mathematical Statistics (collection of papers published 1920-1943)*, Wiley, New York, NY, 1950.

[10] Fisher, R.A., *Statistical Methods and Scientific Inference*, Hafner Press, New York, NY, 3rd ed., 1973.

[11] Stein, S.K., *Calculus and Analytic Geometry*, McGraw-Hill Book Company, New York, NY, 3rd ed., 1982.

[12] Crassidis, J.L. and Markley, F.L., "New Algorithm for Attitude Determination Using Global Positioning System Signals," *Journal of Guidance, Control, and Dynamics*, Vol. 20, No. 5, Sept.-Oct. 1997, pp. 891–896.

[13] Bard, Y., *Nonlinear Parameter Estimation*, Academic Press, New York, NY, 1974.

[14] Walter, E. and Pronzato, L., *Identification of Parametric Models from Experimental Data*, Springer Press, Paris, France, 1994.

[15] Schoukens, J. and Pintelon, R., *Identification of Linear Systems, A Practical Guide to Accurate Modeling*, Pergamon Press, Oxford, Great Britain, 1991.

[16] Horn, R.A. and Johnson, C.R., *Matrix Analysis*, Cambridge University Press, Cambridge, MA, 1985.

[17] Toutenburg, H., *Prior Information in Linear Models*, John Wiley & Sons, New York, NY, 1982.

[18] Magnus, J.R., *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, New York, NY, 1997.

[19] Hoerl, A.E. and Kennard, R.W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, Vol. 12, No. 1, Feb. 1970, pp. 55–67.

[20] Vinod, H.D., "A Survey of Ridge Regression and Related Techniques for Improvements Over Ordinary Least Squares," *The Review of Economics and Statistics*, Vol. 60, No. 1, Feb. 1978, pp. 121–131.

[21] Golub, G.H. and Van Loan, C.F., "An Analysis of the Total Least Squares Problem," *SIAM Journal on Numerical Analysis*, Vol. 17, No. 6, Dec. 1980, pp. 883–893.

[22] Huffel, S.V. and Vandewalle, J., "On the Accuracy of Total Least Squares and Least Squares Techniques in the Presence of Errors on all Data," *Automatica*, Vol. 25, No. 5, Sept. 1989, pp. 765–769.

[23] Björck, Å., *Numerical Methods for Least Squares Problems*, Society for Industial and Applied Mathematics, Philadelphia, PA, 1996.

[24] Huffel, S.V. and Vandewalle, J., *The Total Least Squares Problem: Computational Aspects and Analysis*, Society for Industial and Applied Mathematics, Philadelphia, PA, 1991.

[25] Gleser, L.J., "Estimation in a Multivariate Errors-in-Variables Regression Model: Large Sample Results," *Annals of Statistics*, Vol. 9, No. 1, Jan. 1981, pp. 24–44.

[26] Golub, G.H., Heath, M., and Wahba, G., "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, Vol. 21, No. 2, May 1979, pp. 215–223.