Chapter 3 Estimation of Random Variables

Contents

3 Estimation of Random Variables

3.1 Estimation of Variables

In the sciences, especially engineering and physics, we are required to estimate variables that are not directly observable but are observed only through some other measurable variables – for example, estimating the amplitude, frequency, or phase of a known signal in noise.

3.1.1 Basic Formulation for Estimation of Random Variables

Say we wish to estimate a random variable X by observing another random variable Z that is statistically related to X. An **estimator** for X, call it $\hat{X}$, is some function, $g(\cdot)$, of the observable Z and is written as

$$\hat{X} = g(Z) \tag{3.1}$$

Thus $\hat{X}$ is itself a random variable. The function $g(Z)$ could be an unrestricted nonlinear form, or in special cases could be restricted to be of a particular form like a constant, linear, or polynomial function of the observed variable Z. The **estimation error** is defined as the difference between the true value X and the estimated value as follows:

$$error \overset{\Delta}{=} \varepsilon = (true \quad value) - (estimated \quad value) = X - \hat{X} \tag{3.2}$$

This error is a random variable and is a measure of how well our estimator is performing. It seems reasonable that we would like this error to have a zero mean and a very small variance $E\left[\left(X-\hat{X}\right)^2\right]$.

3.1.2 Bayes Performance Measure

A **performance measure** J, or "yardstick" by which we judge the **optimality** of our estimate, can take many function form. One of most useful is described by the expected value of a **cost function** $C(\cdot)$ that is a function of error or difference between the true value and estimated value. This is called the **Bayes performance measure**, and it can be expressed as

$$J \overset{\Delta}{=} E[C(\varepsilon)] \quad \text{where} \quad \varepsilon = X - \hat{X} \tag{3.3}$$

The **Bayes estimate** of X in term of Z is given by $\hat{X} = g(Z)$, where $g(Z)$ is selected such

that the J, given above, is minimized. If $C(\varepsilon)$ is selected as ε^2, then the performance measure is called the **mean squared error** and $g(Z)$ is the **minimum mean squared error estimate** of X. The mean squared error performance measure will be shown to be mathematically tractable for many situations involving estimating random variables in terms of other random variables.

Other common choices for $g(x)$ include the average absolute error and the average uniform error which are shown along with the average squared error in Figure 3.1. The mean squared error usually requires less statistical information and is more mathematically tractable than the others. Thus our study will concentrate in the beginning on using the mean squared error performance criterion. Estimates using the absolute error and the uniform cost functions will be presented in Section 3.5.

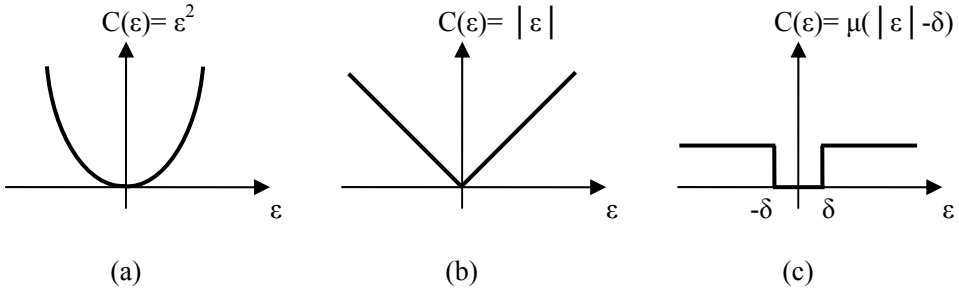

Figure 3.1 Common performance measures
(a) squared error, (b) absolute error, (c) uniform error

3.1.3 Statistical Characterizations of Data

It will be shown that for linear estimation the second-order characterizations of the X and the Z random variables, including means, correlations, and variances, are sufficient to find the minimum mean squared error estimates. Other cost functions could require more statistical properties of X and Z, including higher-order moments and higher-order joint moments for X and Z. For the case of the absolute and uniform error cost functions, a total characterization of X and Z involving the joint and conditional density functions may be required. Thus, in general, more statistical information is usually required for obtaining optimum nonlinear estimators than for optimal linear estimators. Our discussion begins with finding optimal linear estimates followed by optimal nonlinear estimates.

3.2 Linear MMSE Estimation

3.2.1 Estimation of a Random Variable by a Constant

In many cases we wish to approximate a random variable by a deterministic constant. It may be desirable to choose this constant in such a way that a defined performance index is either maximized or minimized. One such performance index is the mean squares error.

Given a random variable Y that we desire to approximate by a constant a such that the mean squared error e_{ms} given by

$$e_{ms} = E\left[(Y-a)^2\right] = \int_{-\infty}^{\infty} (y-a)^2 f_Y(y)dy \tag{3.4}$$

is minimized. This selection of the optimum a corresponds to the best estimate of Y by a constant. A necessary condition for the minimization of e_{ms} can be obtained by taking the derivative of e_{ms} with respect to a and setting the result equal to zero to give

$$\frac{de}{da} = \int_{-\infty}^{\infty} 2(y-a)(-1) f_Y(y)dy = 0 \tag{3.5}$$

The interchange of the derivative and expected value, which is also an integral, is subject to the same rules for interchanging derivatives and integrals. Solving the equation above for a results in

$$a = \int_{-\infty}^{\infty} y f_Y(y)dy = E[y] \tag{3.6}$$

Clearly, it is seen that the constant a that minimizes the mean square error in approximating a random variable by a constant is the mean of the random variable. The corresponding minimum mean squared error using this approximation can be obtained by substituting (3.6) into (3.4), and the variance of Y as follows:

$$e_{mms} = \int_{-\infty}^{\infty} (y-E[y])^2 f_Y(y)dy = {\sigma_Y}^2 \tag{3.7}$$

Example 3.1

Given a random variable Y with probability density function $f_Y(y)$ as $f_Y(y) = be^{-by}\mu(y)$. Find the best constant α that approximates the random variable Y, where best is in the sense that the mean squared error is minimized. Determine the minimum mean squared error for the approximation.

Solution

From Eq.(3.6) the best estimate $\hat{Y} = a$, in the minimum mean squared error sense, is the mean or expected value of Y given by

$$a = \int_{-\infty}^{\infty} ybe^{-by}\mu(y)dy = \frac{e^{-by}}{-b}\left(y+\frac{1}{b}\right)\Big|_0^{\infty} = \frac{1}{b}$$

From Eq.(3.4) or (3.7) the minimum mean squared error is

$$e_{mms} = \int_0^\infty \left(y - \frac{1}{b} \right)^2 be^{-by} dy$$
$$= \sigma_Y^2 = \frac{1}{b^2}$$

3.2.2 Linear estimation of One Random Variable from Another

In other problems we may have access to one random variable X and desire to approximate or estimate another random variable Y that is somehow related to X. Assume that the two random variables X and Y have known means, standard deviations, and correlation specified by η_X, η_Y, σ_X, σ_Y, and R_{XY}, respectively. Let $\hat{Y}$ be the approximation of X and fix it as a linear function of X as

$$\hat{Y} = aX + b \tag{3.8}$$

It is desired to select the a and b such that the mean squared error e_{ms} given by

$$e_{ms} = E\left[\left(Y - \hat{Y} \right)^2 \right] = E\left[\left(Y - \left(aX + b \right) \right)^2 \right] \tag{3.9}$$

is minimized. Necessary conditions for a and b to minimize e_{ms} are

$$\frac{\partial e_{ms}}{\partial a} = E\left[2\left(Y - \left(aX + b \right) \right)\left(-X \right) \right] = 0 \tag{3.10}$$

$$\frac{\partial e_{ms}}{\partial b} = E\left[2\left(Y - \left(aX + b \right) \right)\left(-1 \right) \right] = 0 \tag{3.11}$$

These equations can be simplified by taking expected value and rearranging to yield

$$\begin{aligned} E\left[XY \right] &= aE\left[X^2 \right] + b\eta_X \\ \eta_X &= a\eta_X + b \end{aligned} \tag{3.12}$$

Solving the second equation for b, substituting the result into the first equation, and solving for a gives a and b as

$$\begin{aligned} a &= \frac{E\left[XY \right] - \eta_X \eta_Y}{E\left[X^2 \right] - \eta_X{}^2} = \frac{R_{XY} - \eta_X \eta_Y}{\sigma_X{}^2} = \frac{\mu_{XY}}{\sigma_X{}^2} \\ b &= \eta_Y - a\eta_X = \eta_Y - \frac{\mu_{XY}}{\sigma_X^2}\eta_X \end{aligned} \tag{3.13}$$

Since the matrix of the second derivatives, the Hessian, can be shown to be positive definite, this a and b will give the minimum mean square error. The corresponding minimum mean squared error is easily determined by substituting the a and b given in Eq. (3.13) into Eq. (3.9) and simplifying to give

$$e_{mms} = \sigma_Y{}^2 - \frac{\mu_{XY}^2}{\sigma_X^2} \tag{3.14}$$

Result (3.14) could also have been determined by expanding out (3.9) as follow:

$$e_{ms} = E\left[\left(Y-\left(aX+b\right)\right)Y\right] - E\left[\left(Y-\left(aX+b\right)\right)\left(aX+b\right)\right] \tag{3.15}$$

From Eqs. (3.10) and (3.11) the second expected value is zero, and after replacing the b in the first term by $\eta_Y - a\eta_X$ and writing the second Y as $Y-\eta_Y+\eta_Y$, we write the minimum mean squared error as

$$e_{ms} = E\left(\left(Y-\eta_Y-a\left(X-\eta_X\right)\right)\left(Y-\eta_Y+\eta_Y\right)\right) \tag{3.16}$$

Using the a from Eq. (3.13) and taking the expected value, we easily see the equation above for e_{ms} to be the same as Eq. (3.14).

The minimum mean squared error is seen to be the original variance of Y, the random variable that is approximated, reduced by the ratio of the covariance between X and Y squared and the variance of X. If the covariance is zero, then $a=0$, and the minimum mean squared error is equal to the variance of Y, which implies that we gain no information about Y from X by using a linear function.

Notice that the result depends only on the second-order moments of the involved random variables and that the joint density is not needed. This means that the results do not depend on the probability density functions, and thus all random variables that have the same second-order moments give the same results. If the second-order moments are not given but the joint probability density is the specified information, it is a trivial matter to compute the necessary moments, as shown in the following example.

Example 3.2

Random variables X and Y are characterized by their joint probability density function $f_{XY}(x,y)$, written as

$$f_{XY}(x,y) = \begin{cases} 2, & \{(x,y): 0 \le y \le x, 0 \le x \le 1\} = R_1 \\ 0, & elsewhere \end{cases}$$

The region R_1 is shown in Figure 3.2. Find the best linear estimate of Y in terms of X, where best is in the minimum mean squared error sense and determine the corresponding minimum mean squared error.

Solution

To obtain the best estimator of Y in terms of X, it is necessary to compute the second-order moments, which include all means, variances, and cross-correlation. These moments are calculated as follows: First, the means are determined as

$$E[X] = \int_0^1\int_0^x x f_{XY}(x,y)dxdy = \int_0^1 (2xy]_0^x dx = \frac{2}{3}$$

$$E[Y] = \int_0^1\int_0^x y f_{XY}(x,y)dxdy = \int_0^1 (y^2]_0^x dx = \frac{1}{3}$$

Next, the individual second-order moments are determined as

$$E[X^2]=\int_0^1\int_0^x x^2 f_{XY}(x,y)dxdy=\int_0^1\left(2x^2y\right]_0^x dx=\frac{2}{3}$$

$$E[Y^2]=\int_0^1\int_0^x y^2 f_{XY}(x,y)dxdy=\int_0^1\left(y^3/3\right]_0^x dx=\frac{1}{12}$$

$$E[XY]=\int_0^1\int_0^x xyf_{XY}(x,y)dxdy=\int_0^1\left(xy^2\right]_0^x dx=\frac{1}{4}$$

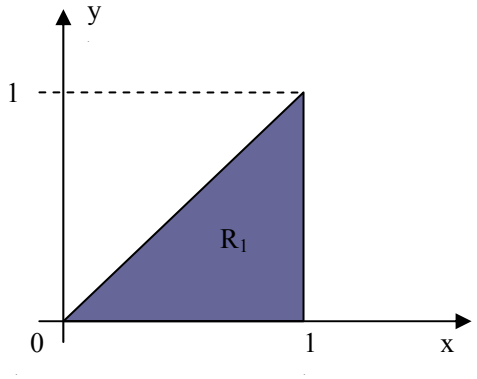

Figure 3.2 R_1 for Example 3.2.

From these moments the individual variances and the covariance are determined, by definition, as

$$\sigma_X^2=E[X^2]-E^2[X]=\frac{1}{2}-\left(\frac{2}{3}\right)^2=\frac{1}{18}$$

$$\sigma_Y^2=E[Y^2]-E^2[Y]=\frac{1}{6}-\left(\frac{1}{3}\right)^2=\frac{1}{18}$$

$$\mu_{XY}=E[XY]-E[X]E[Y]=\frac{1}{4}-\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)=\frac{1}{36}$$

Using Eq.(3.13), the a and b for the minimum mean squared error linear estimate of Y in terms of X are determined from the various first- and second-order moments as

$$a=\frac{E[XY]-E[X]E[Y]}{E[X^2]-E^2[X]}=\frac{\frac{1}{4}-\left(\frac{2}{3}\times\frac{1}{3}\right)}{\frac{1}{18}}=\frac{1}{2}$$

$$b=E[Y]-aE[X]=\frac{1}{3}-\frac{1}{2}\times\frac{2}{3}=0$$

Thus we have the best estimate of Y in terms of X as $\hat{Y}=X/2$.

The minimum mean squared error can be determined from Eq.(3.14) as

$$e_{mms} = \sigma_Y^2 - \frac{\mu_{XY}^2}{\sigma_X^2} = \frac{1}{18} - \frac{\left(\frac{1}{36}\right)^2}{\frac{1}{18}}$$

$$= \frac{1}{18} - \frac{1}{72} = \frac{1}{24}$$

The first term $1/18$, being the variance, is the error for best estimate of Y in terms of the constant—namely the mean—while the second term is the reduction in the total mean squared error that results from the linear estimation thus giving the final minimum mean squared error for the linear estimate as the $1/24$.

3.2.3 Linear Estimation of a Random Variable from N Random Variables

The linear estimation problem can be easily extended to the case of estimating the random variable X_0 in terms of X_1, X_2, … , X_n. It is assumed that the statistical relationships between the random variables are characterized by specifying the means and covariance (a second-order characterization):

$$\begin{aligned} &E[X_i] = \eta_i && i = 0,1,\ldots,n \\ &E\left[(X_i - \eta)_i (X_j - \eta_j)\right] = \mu_{ij}, && i,j = 0,1,\ldots,n \end{aligned} \tag{3.17}$$

The linear estimate of X_0 in terms of X_1, X_2, … , X_n and the mean squared error are given by

$$\hat{X}_0 = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b \tag{3.18}$$

$$e_{ms} = E\left[\left(X_0 - \hat{X}_0\right)^2\right] = E\left[\left(X_0 - \left(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b\right)\right)^2\right] \tag{3.19}$$

The problem then is one of selecting a_1, a_2, … , a_n, and b such that e_{ms} is minimized. Taking partial derivatives of e_{ms} with respect to the a_i and b, setting them equal to zero, and rearranging gives the following set of $n+1$ simultaneous linear equations in $n+1$ unknowns:

$$\begin{aligned} E[X_0 X_1] &= a_1 E[X_1^2] + a_2 E[X_1 X_2] + \cdots + a_n E[X_1 X_n] + b E[X_1] \\ E[X_0 X_2] &= a_1 E[X_2 X_1] + a_2 E[X_2^2] + \cdots + a_n E[X_2 X_n] + b E[X_2] \\ &\vdots \\ E[X_0 X_n] &= a_1 E[X_n X_1] + a_2 E[X_n X_2] + \cdots + a_n E[X_n^2] + b E[X_n] \\ E[X_0] &= a_1 E[X_1] + a_2 E[X_2] + \cdots + a_n E[X_n] + b \end{aligned} \tag{3.20}$$

Solving the last equation for b, substituting the results into the other n equations, rearranging, and using $\mu_{ij} = E[X_i X_j] - \eta_i \eta_j$ results in the following set of n linear equations in n unknowns:

$$\begin{aligned}\mu_{01} &= a_1\mu_{11} + a_2\mu_{12} + \cdots + a_n\mu_{1n}\\ \mu_{02} &= a_1\mu_{21} + a_2\mu_{22} + \cdots + a_n\mu_{2n}\\ &\vdots\\ \mu_{0n} &= a_1\mu_{1n} + a_2\mu_{2n} + \cdots + a_n\mu_{nn}\end{aligned} \tag{3.21}$$

The solution of Eq. (3.21) gives the a_i that minimize the mean square error and the last equation in (3.20) gives the corresponding b in terms of the a_i and the given means as

$$\begin{aligned}b &= E[X_0] - a_1E[X_1] - a_2E[X_2] - \cdots - a_nE[X_n]\\ &= \eta_{X_0} - a_1\eta_{X_1} - a_2\eta_{X_2} - \cdots - a_n\eta_{X_n}\end{aligned} \tag{3.22}$$

After substituting the optimal a_i from the solution of (3.21) and optimal b from (3.22) into the mean squared error of (3.19) and simplifying, the minimum mean squared error can be shown to be

$$e_{mms} = \sigma_{X_0}^2 - \sum_{i-1}^{n}\mu_{0i}a_i \tag{3.23}$$

The proof for this result will be presented after the introduction of the orthogonality principle by Eqs. (3.30) through (3.34).

The Orthogonality Principle.

The problem of linear minimum mean squared error estimation will be viewed from a slightly different perspective. We have $n+1$ random variables X_0, X_1, X_2, … , X_n whose means and second order moments are given by:

$$\begin{aligned}&E[X_i] = \eta_i && i = 0,1,\ldots,n\\ &E[(X_i - \eta)_i(X_j - \eta_j)] = \mu_{ij}, && i,j = 0,1,\ldots,n\end{aligned} \tag{3.24}$$

The random variable X_0 is desired to be approximated by a linear combination of the so-called **data** random variables X_1, X_2, … , X_n as follows:

$$\hat{X}_0 = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b \tag{3.25}$$

The error e involved with this approximation is given by

$$\xi = X_0 - \hat{X}_0 = X_0 - (a_1X_1 + a_2X_2 + \cdots + a_nX_n + b) \tag{3.26}$$

As in the preceding section we desire the coefficients a_1, a_2, … , a_n and b such that the means squared error given below is minimized:

$$e_{ms} = E[\xi^2] = E[(X_0 - (a_1X_1 + a_2X_2 + \cdots + a_nX_n + b))^2] \tag{3.27}$$

By taking the partials of e_{ms} with respect to the a_i and b, it can be easily shown that

$$E[(X_0 - (a_1X_1 + a_2X_2 + \cdots + a_nX_n + b))X_i] = 0, \qquad i = 1,2,\ldots,n \tag{3.28}$$

and

$$E[X_0] - a_1E[X_1] - a_2E[X_2] - \cdots - a_nE[X_n] - b = 0$$

or

$$b=\eta_0-\left(a_1\eta_1+a_2\eta_2+\cdots+a_n\eta_n\right) \tag{3.29}$$

Since the $X_i, i=1,2,\ldots,n$ are the **data** and $X_0-\left(a_1X_1+a_2X_2+\cdots+a_nX_n+b\right)$ is the **error**, Eq. (3.28) can be interpreted as the **error is orthogonal to each member of the data**, and (3.29) gives the solution for the optimal b. This statement is **the first part of the orthogonality principle** for minimum mean squared error linear estimation. The second part of the orthogonality principle deals with the evaluation of the minimum mean squared error for this optimal estimator. The e_{mms} of Eq. (3.27), where the a_i are the solution to the simultaneous equations of (3.28), can be written in the expanded form as

$$\begin{aligned} e_{mms}&=E\left[\left(X_0-\left(a_1X_1+a_2X_2+\cdots+a_nX_n+b\right)\right)\left(X_0-b\right)\right]\\ &\quad-E\left[\left(X_0-\left(a_1X_1+a_2X_2+\cdots+a_nX_n+b\right)\right)\left(a_1X_1+a_2X_2+\cdots+a_nX_n\right)\right]\end{aligned} \tag{3.30}$$

The second expected value of (3.30) is zero from (3.28), since it is the expected value of the error times a linear combination of the data. Thus the formula for the minimum mean squared error can be reduced to

$$e_{mms}=E\left[\left(X_0-\left(a_1X_1+a_2X_2+\cdots+a_nX_n+b\right)\right)\left(X_0-b\right)\right] \tag{3.31}$$

Eq.(3.31) is **the second part of the orthogonality principle**, and it gives a reduced expression for determining the minimum mean squared error. This result can be written in another form by substituting the b of Eq. (3.29) into (3.31) to get

$$\begin{aligned} e_{mms}=E\{\ &[(X_0-\eta_0)-a_1(X_1-\eta_1)-a_2(X_2-\eta_2)-\cdots-a_n(X_n-\eta_n)]\\ &\cdot[(X_0-\eta_0)-(a_1\eta_1+a_2\eta_2+\cdots+a_n\eta_n)]\ \}\end{aligned} \tag{3.32}$$

The equation above for e_{mms} can be simplified further by multiplying out the last terms in the two parts shown to give

$$\begin{aligned} e_{mms}&=E\{[(X_0-\eta_0)-a_1(X_1-\eta_1)-a_2(X_2-\eta_2)-\cdots-a_n(X_n-\eta_n)](X_0-\eta_0)\}\\ &\quad-E\{[(X_0-\eta_0)-a_1(X_1-\eta_1)-a_2(X_2-\eta_2)-\cdots-a_n(X_n-\eta_n)](a_1\eta_1+a_2\eta_2+\cdots+a_n\eta_n)\}\end{aligned} \tag{3.33}$$

The first expected value give a sum of covariance of X_0 and X_i, and the second expected value is zero, since it gives a weighted sum of central moments of the X_i, $E[X_i-\eta_i]$, which are zero by definition for all $i=0,1,2,\ldots,n$. Thus the expression above for the minimum mean squared error can be written in term of just the covariance between X_0 and the data set X_i and the optimal coefficients as

$$e_{mms}=\sigma_{X_0}^2-\sum_{i=1}^{n}\mu_{0i}a_i \tag{3.34}$$

Special case. For the many problems where the expected value of the X_i, $i=0,1,\ldots,n$ are all zero the orthogonality principle can be simplified. From Eq. (3.29) it can be shown that $b=0$ and then from (3.28) that the a_i are the solution of the set of equations

$$E\left[\left(X_0-\left(a_1X_1+a_2X_2+\cdots+a_nX_n\right)\right)X_i\right]=0,\qquad i=1,2,\ldots,n \tag{3.35}$$

Setting $b=0$ in (3.31) gives the resulting minimum mean squared error as the expected value of the product of the error and the variable X_0 that we are estimating

$$e_{mms}=E\left[\left(X_0-\left(a_1X_1+a_2X_2+\cdots+a_nX_n\right)\right)X_0\right] \tag{3.36}$$

Multiplying out the terms in (3.36) and taking expected values of the products allows the minimum mean squared error to be written in terms of the correlations between the data as follows:

$$e_{mms}=\sigma_{X_0}^2-\sum_{i=1}^{n}a_iR_{0i} \tag{3.37}$$

The first term of this error, $\sigma_{X_0}^2$, is the error that would occur if X_0 were estimated by a constant with no data available, namely the mean of aero, whereas the terms in the sum represent the reduction in mean squared error due to the use of each member in the data set.

Example 3.4

Suppose that we have the random variables S, Z_1, Z_2,…, Z_n where the Z_i are related to S by

$$Z_i=S+N_i,\qquad i=1,2,...,n$$

Let S be a zero mean random variable with known variance σ_S^2 and the $N_i, i=1,2,...,n$, be zero mean mutually orthogonal random variables each with variance σ_N^2 and each orthogonal to S. These assumptions can be written as using the Kronecker delta function δ_{ij} as

$$E\left[N_iN_j\right]=\sigma_N^2\delta_{ij},\qquad i,j=1,2,...,n$$

$$E\left[SN_i\right]=0,\qquad i=1,2,...,n$$

From the Z_i we wish to approximate or estimate S by a linear combination of the Z_i given by

$$\hat{S}=\sum_{i=1}^{n}a_iZ_i+b$$

Such that the mean squared error is minimized. The problem can be related to estimating a random constant value from a set of n noisy observations.

Solution

The solution is based on a direct application of the orthogonality principle. From Eq.(3.29) the b is as follows:

$$b=E\left[S\right]-a_1E\left[Z_1\right]-a_2E\left[Z_2\right]-\cdots-a_nE\left[Z_n\right]$$

Since $E\left[Z_i\right]=E\left[S\right]+E\left[N_i\right]=0$ for $i=1,2,...,n$ and $E\left[S\right]=0$, it is easily seen that $b=0$. After substituting $b=0$ into the first part of the orthogonality principle, Eq.(3.28), which states that the error is orthogonal to the data, gives

$$E\left[\left(S-\left(a_1Z_1+a_2Z_2+\cdots+a_nZ_n\right)\right)Z_i\right]=0,\qquad i=1,2,...,n$$

This equation could have been obtained directly from the special case of zero means from (3.35). After

taking the expected values, we can rearrange the set of simultaneous equations above to give

$$E[SZ_i] = a_1 E[Z_1 Z_i] + a_2 E[Z_2 Z_i] + \cdots + a_n E[Z_n Z_i], \qquad i = 1,2,...,n$$

To describe this set of equations, it is necessary to find the correlation of S and Z_i for all i and the correlation of Z_i and Z_j for all i and j. First the correlation of S and Z_i is determined as

$$E[SZ_i] = E[S(S+N_i)] = E[S^2] + E[SN_i] = \sigma_S^2, \qquad i = 1,2,...,n$$

The second term in the sum above, $E[SN_i]$, is zero by the original given properties on the expected value of SN_i. Similarly the $E[Z_i Z_j]$ can be found by using the given orthogonal properties of N_i and N_j as

$$\begin{aligned} E[Z_i Z_j] &= E[(S+N_i)(S+N_j)] = E[S^2] + E[N_i S] + E[SN_j] + E[N_i N_j] \\ &= \sigma_S^2 + \sigma_N^2 \delta_{ij} \end{aligned}$$

Substituting these results into the simultaneous equations given above results in the following set of n simultaneous equations in the n unknowns $a_1, a_2, ..., a_n$:

$$\begin{aligned} \sigma_S^2 &= (\sigma_S^2 + \sigma_N^2) a_1 + \sigma_S^2 a_2 + ... + \sigma_S^2 a_n \\ \sigma_S^2 &= \sigma_S^2 a_1 + (\sigma_S^2 + \sigma_N^2) a_2 + ... + \sigma_S^2 a_n \\ &\vdots \\ \sigma_S^2 &= \sigma_S^2 a_1 + \sigma_S^2 a_2 + ... + (\sigma_S^2 + \sigma_N^2) a_n \end{aligned}$$

By symmetry of the equations above, the a_i are equal, so the solution can be obtained from any of the equations above by equating the a_i as follows:

$$\sigma_S^2 = n\sigma_S^2 a_i + \sigma_N^2 a_i$$

Solving for a_i gives

$$a_i = \frac{\sigma_S^2}{n\sigma_S^2 + \sigma_N^2}, \qquad i = 1,2,...,n$$

With these coefficients the approximation of S can be written as

$$\hat{S} = \frac{\sigma_S^2}{n\sigma_S^2 + \sigma_n^2} \sum_{i=1}^{n} Z_i$$

As might be expected, the form of the estimator is a weighted sum of the measurements. This equation can be further rearranged to be written in terms of the average of the measurements in the following ways:

$$\hat{S} = \frac{n\sigma_S^2}{n\sigma_S^2 + \sigma_n^2} \left(\frac{1}{n} \sum_{i=1}^{n} Z_i \right) \quad or$$

$$\hat{S} = \frac{1}{1 + \sigma_n^2 n^{-1} \sigma_S^{-2}} \left(\frac{1}{n} \sum_{i=1}^{n} Z_i \right)$$

Using these a_i in the second part of the orthogonality principle equation (3.34), and identifying X_0 as S and X_i as Z_i, we can write the minimum mean squared error as

$$e_{mms} = \sigma_S^2 - \sum_{i=1}^{n} a_i E[SZ_i]$$

With the correlations of S and Z_i determined as above, e_{mms} becomes

$$e_{mms} = \sigma_S^2 - \sum_{i=1}^{n} \frac{\sigma_S^2}{n\sigma_S^2 + \sigma_N^2}\sigma_S^2$$
$$= \frac{\sigma_N^2 \sigma_S^2}{n\sigma_S^2 + \sigma_N^2}$$

For the special case where n, σ_S^2, and σ_N^2 are such that the effective signal to noise ratio, $n\sigma_S^2/\sigma_N^2 >> 1$, the coefficient for the linear approximation and the corresponding minimum mean squared error become

$$a_i \approx \frac{\sigma_S^2}{n\sigma_S^2} = \frac{1}{n}, \qquad e_{mms} \approx \sigma_N^2 \frac{\sigma_S^2}{n\sigma_S^2} = \frac{\sigma_N^2}{n}$$

The mean squared error is thus seen to be inversely proportional to the number of observations, and the coefficients make the estimate an average of the observations.

$$\hat{S} \approx \frac{1}{n}\sum_{i=0}^{n} Z_i$$

If n and σ_S^2 are such that $n\sigma_S^2 << \sigma_N^2$, low effective signal to noise ratio, the a_i and mean squared error become

$$a_i \approx \frac{\sigma_S^2}{\sigma_N^2}, \qquad e_{mms} \approx \sigma_S^2$$

This shows that in the low effective signal to noise case that the coefficients approach the ratio of the variances of S and N and the minimum mean squared error approaches the variance of S. Thus the form of the estimator is

$$\hat{S} \approx \frac{\sigma_S^2}{\sigma_N^2}\sum_{i=0}^{n} Z_i$$

Since the variance of S is the initial mean squared error without our taking any measurements, the results indicate that as σ_S^2/σ_N^2 approaches zero, the data do not significantly reduce the mean squared error. Therefore for this case the observations are almost not used to estimate S, which for this problem is zero, is used as the best estimate.

Example 3.5

Say we are given a random variable X that is exponentially distributed with probability density function

given by

$$f_X(x)=2e^{-2x}\mu(x)$$

Define a random variable Y related to X through $Y=e^{-X}$. We wish to find the best linear estimate, in the minimum mean squared error sense, of Y in terms of X. Let the optimal estimate be given by $\hat{Y}=aX+b$. Determine the a and b for this optimum estimate, and determine its minimum mean squared error.

Solution

It is known that the optimal linear estimate can be found by using the orthogonality principle, which states that the error is orthogonal to the data. For our problem this relationship can be expressed as

$$E[(Y-aX-b)X]=0$$

Where the first term in the ( ) is the error and X is the data. Expanding out and taking the expected value gives an equation for a and b in terms of various first- and second-order moments of X and Y as

$$E[YX]=aE[X^2]+bE[X]$$

The other equation of the orthogonal principle from Eq.(3.29), recognizing that $X_0=Y$ and $X_1=X$, is

$$b=E[Y]-aE[X]$$

To solve these two equations for a and b we require the calculation of the various moments indicated. These expected values will now be determined

$$E[Y]=E[e^{-X}]=\int_0^\infty 2e^{-3x}dx=\frac{2}{3}$$

$$E[YX]=E[e^{-X}X]=\int_0^\infty 2xe^{-3x}dx=\frac{2}{9}$$

$$E[Y^2]=E[e^{-2X}]=\int_0^\infty 2e^{-4x}dx=\frac{1}{2}$$

$$E[X]=\int_0^\infty 2xe^{-2x}dx=\frac{1}{2}$$

$$E[X^2]=\int_0^\infty 2x^2e^{-2x}dx=\frac{1}{2}$$

Thus with these values the simultaneous equations become

$$\frac{2}{9}=\frac{1}{2}a+\frac{1}{2}b$$

$$b=\frac{1}{3}-\frac{1}{2}a$$

Solving for a and b yields $a=4/9$ and $b=8/9$. Therefore the best estimate is

$$\hat{Y}=-\frac{4}{9}X+\frac{8}{9}$$

From Eq.(3.23) the minimum mean squared error can be written as

$$e_{mms} = \sigma_Y^2 - \mu_{YX} a$$
$$= \left(\frac{1}{2} - \frac{4}{9}\right) - \left(\frac{2}{9} - \frac{1}{3}\right)\left(-\frac{4}{9}\right) = \frac{1}{162}$$

3.3 Nonlinear MMSE Estimation

In many problems a further reduction in mean squared error can be obtained by using nonlinear estimation. This reduction is at the expense of requiring more statistical properties of the random variables. It will now be show that nonlinear minimum mean squared error estimation can be performed if almost a total characterization of the random variables is specified. This is in sharp contrast with linear estimation where only first- and second-order moments are needed, that is, the means, variances, and covariances of the random variables involved.

3.3.1 Nonlinear Estimation of One Random Variable from Another

Given two random variable X and Y that are statistically related with joint probability density function $f_{XY}(x,y)$. By observing X, we desire to estimate Y by some nonlinear function $g(\cdot)$ of X, not necessarily a linear function, as

$$\hat{Y} = g(X) \tag{3.38}$$

Our estimate is to be the best in the sense that it gives the minimum mean squared error e as

$$e_{mms} = E\left[\left(Y - \hat{Y}\right)^2\right] = E\left[\left(Y - g(X)\right)^2\right] \tag{3.38}$$

The first step in minimizing e_{ms} is to write it in terms of $f_{XY}(x,y)$ as

$$e_{ms} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(y - g(x)\right)^2 f_{XY}(x,y)dxdy \tag{3.40}$$

Using the fact that the joint density is a product of the conditional and marginal densities, e_{ms} can be rearranged to yield

$$e_{ms} = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\left(y - g(x)\right)^2 f_Y(y|x)dy\right] f_X(x)dx \tag{3.41}$$

Since the product of $f_Y(y|x)$ and $\left(y - g(x)\right)^2$ is positive for all x and $f_X(x)$ is always positive, e can be minimized by making the term in bracket as small as possible for every x. That is, select $g(\cdot)$ in such a way that the term in brackets, $I(x)$, given by

$$I(x) = \int_{-\infty}^{\infty}\left(y - g(x)\right)^2 f_Y(y|x)dy \tag{3.42}$$

is minimized. But for each x, $g(x)$ is just a constant, and we have shown previously (Section 3.2.1) that the integral is smallest when $g(x)$ is selected as the mean of the random variable, which for this case is in terms of the conditional probability density $f_Y(y|x)$. Therefore the optimum $g(x)$, in the minimum mean squared error sense, is given by

$$g(x)=E[Y|X=x]=\int_{-\infty}^{\infty} y f_Y(y|x)dy \tag{3.43}$$

The resulting $g(x)$ is sometimes called the **nonlinear regression curve**. The minimum mean squared error is obtained by substituting this optimal $g(x)$ into (3.39), the equation for the mean squared error, to give

$$\begin{aligned} e_{mmse} &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}(y-E[Y|x])^2 f_Y(y|x)dy\right] f_X(x)dx \\ &= \int_{-\infty}^{\infty} \sigma_{Y|x}{}^2 f_X(x)dx \end{aligned} \tag{3.44}$$

Thus the minimum mean squared error is seen to be $E\left[(Y-E[Y|X])^2\right]$, where $E[Y|X]$ given by $E[Y|X=x]_{x=X}$ is random variable.

Example 3.6

The density for Y conditioned on $X=x$ and the marginal density $f_X(x)$ of two random variables X and Y are given by

$$f_Y(y\,|\,x)=xe^{-yx}\mu(y), \qquad f_X(x)=\mu(x-1)-\mu(x-2)$$

Find the best nonlinear estimate of Y in terms of X, and calculate the corresponding minimum mean squared error for your estimate.

Solution

The conditional density in y is seen as exponential with parameter $a=x$; thus we are trying to estimate this parameter from the given information and a sample of X. The $g(x)$ that gives the best nonlinear estimate in the minimum mean squared error sense is the conditional mean

$$\begin{aligned} g(x) &= E[Y\,|\,X=x]=\int_{-\infty}^{\infty} y f_y(y\,|\,x)dy \\ &= \int_0^{\infty} yxe^{-xy}dy = \frac{xe^{-xy}(xy-1)}{x^2}\Big]_0^{\infty} = \frac{1}{x} \end{aligned}$$

Therefore $\hat{Y}=g(X)=1/X$.

The minimum mean squred error for the optimal $g(x)$ can be determined from (3.41) as

$$e_{mmse}=\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\left(y-\frac{1}{x}\right)^2 f_Y(y|x)dy\right] f_X(x)dx$$

But the term in brackets is the conditional variance for the exponential density, which from Eq.2.38 is just $1/x^2$. Therefore the mean squared error can be found as follows:

$$e_{mmse} = \int_{-\infty}^{\infty} \frac{1}{x^2} f_X(x)dx = \int_1^2 \frac{1}{x^2} 1dx = \frac{-1}{x}\Big]_1^2 = \frac{1}{2}$$

Example 3.7

The random variable X and Y are statistically characterized by the following joint probability denstity function:

$$f(x,y) = \begin{cases} 2, & (x,y) \in R = \{(x,y): 0 < y < x, 0 < x < 1\} \\ 0, & elsewhere \end{cases}$$

Where R is shown in Figure 3.3.

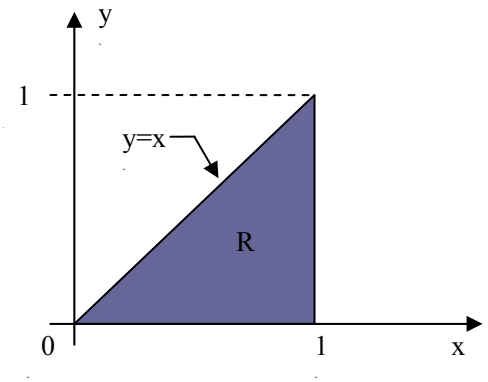

Figure 3.3 Region R for Example 3.7.

(a) Find the best nonlinear estimate of Y in terms of X .

(b) Find the best linear estimate of Y in terms of X .

(c) Comment on your results.

Solution

(a) The best nonlinear estimate is the conditional expected value of Y given $X = x$. To find the conditional density, it is necessary to determine the marginal density $f_X(x)$. It is determined by integrating the joint density for each value of x :

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$

For any x such that $0 < x < 1$ the $f_X(x)$ becomes

$$f_X(x) = \int_0^x 2dy = 2x$$

For x outside the interval $(0,1)$ the probability density function is zero, so the conditional density function can be summarized by using the unit step function as

$$f_Y(y/x)=\frac{f_{XY}(y,x)}{f_X(x)}=\frac{1}{x}(\mu(y)-\mu(y-x))$$

The function for the best nonlinear estimate of Y in terms of X for $0<x<1$ is the conditional mean as follows

$$g(x)=E[Y\mid X=x]=\int_{-\infty}^{\infty}yf_Y(y\mid x)dy$$

$$=\int_0^x y\frac{1}{x}dy=\frac{y^2}{2x}\Big]_0^x=\frac{x}{2}$$

It is seen that for this example the best nonlinear estimate is a linear estimate.

(b) The best linear estimate of Y in term of X by $ax+b$ was found in Example 3.2 to be $\hat{Y}=X/2$, so for this example the best linear estimate is the same as the best nonlinear estimate.

(c) It is rare that the best linear and nonlinear estimates are the same, but they will be shown in Section 3.3.3 to be the same for the special case where X and Y are jointly Guassian.

3.3.2 Nonlinear Estimation of One Random Variable from N Random

The Problem of estimating a random variable X_0 in terms of statistically related random variables $X_1,X_2,...,X_n$, where the random variables are characterized by their joint probability density function $f_x(x_0,x_1\cdots,x_n)$ is a common application in the sciences. In Section 3.2 formulas were given for finding the best linear estimate, while in this section the problem of nonlinear estimation will be examined.

We desire to approximate or estimate X_0 by a nonlinear function $g(X_0,X_1,...,X_n)$ as

$$\hat{X}_0=g(X_1,X_2,...,X_n) \tag{3.45}$$

It is desired to find the $g(X_0,X_1,...,X_n)$ that will minimize the mean squared error given by

$$e_{ms}=E[(X_0-\hat{X}_0)^2]=E[(X_0-g(X_1,X_2,...,X_n))^2] \tag{3.46}$$

By extending the results given for the best nonlinear estimate of one random variable in terms of another, it can be easily shown that the optimal function $g(x_1,x_2,...,x_n)$ in the minimum mean squared error sense, is the conditional given by

$$\begin{aligned}g(x_1,x_2,...,x_n)&=E[X_0|X_1=x_1,X_2=x_2,\cdots,X_n=x_n]\\&=\int_{-\infty}^{\infty}x_0f(x_0|x_1,x_2,\cdots,x_n)dx_0\end{aligned} \tag{3.47}$$

It is worth noting that to obtain the minimum mean squared error nonlinear estimate of one random variable in terms of other random variables we require the knowledge of the conditional probability density function, $f(x_0|x_1,x_2,\cdots,x_n)$, which in many problems involves more statistical

information than is available. It was shown in Section 3.2 that restricting $g(\cdot)$ to a linear function reduces the required information to the knowledge of the second-order properties, that is, the means, variances and covariances of the random variables involved.

Example 3.10 (Random Variables)

Let X_0, X_1, and X_2 be random variables. The conditional density for X_0, given $X_1 = x_1$, and $X_2 = x_2$, is known to be the following:

$$f_{X_0}(x_0 \mid x_1, x_2) = k \cdot x_0^{(x_1-1)}(1-x_0)^{x_2-1}, \qquad 0 < x_0 < 1$$

$$where \qquad k = \frac{\Gamma(x_1)}{\Gamma(x_1)\Gamma(x_2)}$$

Find the best nonlinear estimate of X_0 in terms of X_1 and X_2 where best is in the minimum mean squared error sense.

Solution

From Eq.(3.47) we know that the optimal function is the conditional mean. For the conditional density function above the conditional mean can be shown from Eq.(2.50) to be

$$g(x_1, x_2) = E[X_0 \mid x_1, x_2] = \int_0^1 x_0 f_{X_0}(x_0 \mid x_1, x_2) dx_0$$

$$= \frac{x_1}{x_1 + x_2}$$

Thus the best nonlinear estimate of X_0 in terms of X_1 and X_2 is $\hat{X}_0 = X_1/(X_1 + X_2)$.

3.3.3 Nonlinear Estimation of Gaussian Random Variables

Given X and Y are jointly Gaussian random variables with known means, η_X, η_Y, variances, ${\sigma_X}^2$, ${\sigma_Y}^2$, and covariance σ_{XY}. The best nonlinear estimate of Y in term of X is desired where best is in the sense of minimizing the mean squared error. The solution for $g(x)$ has already been shown to be the conditional mean

$$g(x) = E[Y|X = x] = \int_{-\infty}^{\infty} y f_Y(y|x) dy \tag{3.48}$$

From Section 2.6.4, the conditional density for jointly Gaussian random variables is known to be Gaussian with mean and variance given by

$$\eta_{Y|x} = \eta_Y + \frac{\rho_{XY}\sigma_Y(x - \eta_X)}{\sigma_X}$$

$$\sigma^2{}_{Y|x} = {\sigma_Y}^2(1 - \rho^2{}_{XY}) \tag{3.49}$$

Thus the optimal $g(x)$ for this Gaussian case is the conditional mean

$$g(x)=\eta_X+\frac{\rho_{XY}\sigma_Y}{\sigma_X}(x-\eta_X) \tag{3.50}$$

By the definition for the correlation coefficient, Eq. (2.101), where ρ_{XY} is

$$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y} \tag{3.51}$$

the optimal $g(x)$ can be rewritten in terms of the various variances and covariances as

$$\begin{aligned} g(x)&=\eta_X+\frac{\sigma_{XY}}{\sigma^2{}_X}(x-\eta_X)\\ &=\frac{\sigma_{XY}}{\sigma^2{}_X}x+\eta_x\left(1-\frac{\sigma_{XY}}{\sigma^2{}_X}\right) \end{aligned} \tag{3.52}$$

Thus the **best nonlinear estimate** for the jointly **Gaussian** case is seen to be a **linear estimate** given by

$$\begin{aligned} &\hat{Y}=g(X)=aX+b\\ &\textit{where}\qquad a=\frac{\sigma_{XY}}{\sigma_X{}^2},\qquad b=\eta_x\left(1-\frac{\sigma_{XY}}{\sigma_X{}^2}\right) \end{aligned} \tag{3.53}$$

This fact gives more power to the linear estimate for the case of Gaussian random variables and releases us from needing to look for a best nonlinear estimate for that case. However, it is not possible to say that if the best nonlinear estimate is a linear estimate, the random variables involved are jointly Gaussian. A counterexample has already been shown in Example 3.7.

For the special case of zero mean Gaussian random variables, the a and b from Eq. (3.53) reduce to

$$a=\frac{\sigma_{XY}}{\sigma_X^2}\quad\text{and}\quad b=0 \tag{3.54}$$

The extension of the result above to $n+1$ jointly random variables is possible and the best nonlinear estimate of the Gaussian random variable X_0 in terms of the other jointly Gaussian random variables X_1, X_2, … , X_n can be shown to be a linear function of the random variables X_1, X_2, … , X_n. This is expressed as follows:

$$\begin{aligned} g(x_1,\ldots,x_n)&=E[X_0|X_1=x_1,X_2=x_2,\cdots,X_n=x_n]\overset{\Delta}{=}\eta_{X_0|x_1,x_2,\ldots,x_n}\\ &=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}x_0\,f(x_1,x_2,\ldots,x_n|x_0)dx_0\\ &=a_1x_1+a_2x_2+\cdots+a_nx_n+b \end{aligned} \tag{3.55}$$

where the $\mathbf{a}=[a_1,a_2,\ldots,a_n]^T$ and b can be determined from (3.21) and (3.22). If the random vector $\mathbf{X}_{n+1}=[X_0,X_1,\ldots,X_n]^T=[X_0\vdots\mathbf{X}_n]^T$ is jointly Gaussian with given covariance matrix $\mathbf{K}_{n+1}$ and mean vector $\mathbf{m}_{n+1}=[E[X_0],E[X_1],\ldots,E[X_n]]^T=[E[X_0]\vdots\mathbf{m}_n]^T$ then the best nonlinear estimate of

X_0 in terms of X_1, X_2, … , X_n can be written as

$$\hat{X}_0 = [a_1, a_2, ..., a_n] \cdot \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \mathbf{a}\mathbf{X}_n + b \tag{3.56}$$

To obtain the $\mathbf{a}$ vector and b in terms of the given information, the covariance matrix $\mathbf{K}_{n+1}$ is first partitioned as

$$\mathbf{K}_{n+1} = \begin{bmatrix} \sigma_{X_0}^2 & \mu_{X_0}^T \\ \mu_{X_0} & \mathbf{K}_N \end{bmatrix} \tag{3.57}$$

$$where \quad \boldsymbol{\mu}_{X_0}^T = [\mu_{X_0X_1}, \mu_{X_0X_2}, ..., \mu_{X_0X_N}]$$

The desired $\mathbf{a}$ and b can now be obtained by using the linear estimate formulas given in (3.21) and (3.22) as follows

$$\boldsymbol{\mu}_{X_0}^T = \mathbf{K}_n \mathbf{a}, \qquad b = E[X_0] - \mathbf{a}^T \mathbf{m}_n \tag{3.58}$$

where the solution for the coefficient vector $\mathbf{a}$ is easily obtained from the equation above as

$$\mathbf{a} = \mathbf{K}_n^{-1} \boldsymbol{\mu}_{X_0}^T \tag{3.59}$$

3.4 Properties of Estimators of Random Variables

There are a number of properties of estimators of random variables, whether they be linear or nonlinear functions of the observations, that are useful to define. Assume that X is a random variable that is statistically related to another random variable Z. Let $\hat{X}(Z) = g(Z)$ be an estimator for X. Assume that mean $E[X]$ is known or can be determined.

An estimator of a random variable X is **unbiased** if its expected value is equal to the expected value of the random variable being estimated, which can be written as

$$E[\hat{X}(Z)] = E[X] \tag{3.60}$$

If (3.60) is not satisfied, the estimator is **biased**.

Suppose that we are estimating a random variable X from a vector of observations $\mathbf{Z} = [Z_1, Z_2, ..., Z_N]^T$. The **Cramer-Rao bound**, is the most important lower bound on the variable X and is given below without proof:

$$E[(X - \hat{X}(\mathbf{Z}))^2] \geq \frac{1}{E\left[\left(\frac{\partial \ln p(\mathbf{z}|x)}{\partial x}\right)^2\right]} = \frac{1}{-E\left[\frac{\partial^2 \ln p(\mathbf{z}|x)}{\partial^2 x}\right]} \tag{3.61}$$

Any estimator satisfying this bound is called an **efficient estimator**.

3.5 Bayes Estimation

The **performance measure** J_{BAYES}, or "yardstick" by which we judge the **optimality** for a Bayes estimate, is the expected value of a **cost function** $C(\cdot)$ that is a function of the error or difference between the true value and the estimated value.

$$\begin{aligned} J_{BAYES} &\overset{\Delta}{=} E[C(\varepsilon)] \\ where \qquad \varepsilon &= \mathrm{X} - \hat{\mathrm{X}} \end{aligned} \tag{3.62}$$

The Bayes estimator of X in term of Z is given by $\hat{X} = g(Z)$ where $g(Z)$ is selected such that the J_{BAYES}, given above, is minimized. Common choices for $C(\varepsilon)$, shown in Figure 3.1, are the squared error, absolute error, and the uniform error, which are defined as

$$\begin{aligned} C_{MS}(\varepsilon) &= \varepsilon^2 \\ C_{ABS}(\varepsilon) &= |\varepsilon| \\ C_{UN}(\varepsilon) &= \begin{cases} 0, & |\varepsilon| \le \delta \\ 1, & |\varepsilon| > \delta \end{cases} \end{aligned} \tag{3.63}$$

The choice of which cost function to use is subjective. It is seen that the $C_{MS}(\varepsilon)$ penalizes large errors severely and rewards errors less than 1, whereas the $C_{ABS}(\varepsilon)$ treats errors on a linear basis. $C_{ABS}(\varepsilon)$ does not penalize if the absolute value of the error is within the bound of δ but penalizes all errors bigger than δ in magnitude by the same value. Bayes estimates for each of these three cost functions are now presented.

3.5.1 Bayes Estimates

Squared Error Cost. Bayes estimates for random variables X in terms of a measured random variable Z for the squared error cost function have been thoroughly explored in the previous two sections. It was shown that the $g(z)$ for the optimal nonlinear estimate of X was obtained as the conditional mean of X given $Z = z$ and could be written as

$$\begin{aligned} g(z) &= E[X|Z = z] = \int_{-\infty}^{\infty} x p(x|z) dx \\ \hat{X} &= g(z) \end{aligned} \tag{3.64}$$

Thus the Bayes estimate for the squared error cost is seen to be the mean of the posteriori density $p(x|z)$

given by

$$\hat{x}_{MS} = \int_{-\infty}^{\infty} x p(x|z) dx \tag{3.65}$$

Unfortunately, what we usually have available, or can easily obtain, is $p(x|z)$. Thus $p(x|z)$ must be found by using Bayes's rule as follows:

$$p(x|z) = \frac{p(z|x)p(x)}{\int_{-\infty}^{\infty} p(z|x)p(x)dx} \tag{3.66}$$

Absolute Error Cost. To obtain the optimal estimate of X in terms of Z for the absolute error cost, we must select $g(Z)$ such that J_{ABS} given below is minimized:

$$\begin{aligned} J_{ABS} &\stackrel{\Delta}{=} E[C_{ABS}(\varepsilon)] = E\left[\left|X - \hat{X}\right|\right] = E\left[\left|X - g(Z)\right|\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |x - g(z)| p(x,z) dx dz \end{aligned} \tag{3.67}$$

Equation (3.67) can be rearranged by using the relationship for the joint and conditional densities as follows:

$$J_{ABS} = \int_{-\infty}^{\infty} p(z)\left[\int_{-\infty}^{\infty} |x - g(z)| p(x|z) dx\right] dz \tag{3.68}$$

To minimize J_{ABS}, it is sufficient to minimize the term in brackets for every value of z. This is true since $p(z)$ is always positive as is the term in brackets, I, which can be written as

$$I = \int_{-\infty}^{g(z)} (-x + g(z)) p(x|z) dx + \int_{g(z)}^{\infty} (x - g(z)) p(x|z) dx \tag{3.69}$$

For a fixed z, $g(z)$ is a constant setting the partial derivative of the integral with respect to $g(z)$ equal to zero, and using Liebnitz's rule, we obtain a necessary condition for minimizing I as follows:

$$\begin{aligned} \frac{\partial I}{\partial g(z)} &= \int_{-\infty}^{g(z)} (-x + g(z)) p(x|z) dx + \int_{g(z)}^{\infty} (x - g(z)) p(x|z) dx \\ &= (-g(z) + g(z)) p(g(z)|z) + \int_{-\infty}^{g(z)} p(x|z) dx - (g(z) - g(z)) p(g(z)|z) - \int_{g(z)}^{-\infty} p(x|z) dx = 0 \end{aligned} \tag{3.70}$$

Recognizing the zero terms and rearranging the $g(z)$ for each z satisfies the equation

$$\int_{-\infty}^{g(z)} p(x|z) dx = \int_{g(z)}^{-\infty} p(x|z) dx \tag{3.71}$$

In this expression it is seen that the optimum $g(z)$, for each z is the **median of the posteriori density** $p(x|z)$. Thus the optimum estimator in the minimum absolute error sense is seen to be the $\hat{x}_{ABS}(z)$, which satisfies

$$\int_{-\infty}^{\hat{x}_{ABS}(z)} p(x|z) dx = \int_{\hat{x}_{ABS}(z)}^{-\infty} p(x|z) dx \tag{3.72}$$

As with the MS estimate, what we usually have available or can easily obtain is $p(x|z)$. Thus $p(x|z)$ must be found by using Bayes's rule as follows:

$$p(x|z)=\frac{p(z|x)p(x)}{\int_{-\infty}^{\infty}p(z|x)p(x)dx} \tag{3.73}$$

Uniform Cost. To obtain optimal estimator of X in terms of Z for the uniform error cost function, the $g(z)$ is selected such that J_{UN} given below is minimized:

$$\begin{aligned} J_{UN}&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}C_{UN}(x-g(z))p(x,z)dxdz \\ &=\int_{-\infty}^{\infty}p(z)\left[\int_{-\infty}^{\infty}C_{UN}(x-g(x))p(x|z)dx\right]dz \end{aligned} \tag{3.74}$$

The optimal estimator is as before, where the $g(z)$ minimizes the term in brackets given by

$$I(z)=\int_{-\infty}^{\infty}C_{UN}(x-g(z))p(x|z)dx=1-\int_{|x-g(z)|<\delta}p(x|z)dx \tag{3.75}$$

Minimizing $I(z)$ for each z is equivalent to maximizing the integral on the right side of Eq. (3.75). If δ is very small, then that integral is maximized by selecting $g(z)$ to be the maximum of $p(x|z)$ viewed as a function of x for each z. So the estimator in the uniform cost sense is written as

$$\hat{x}_{UN}(z)=\max_{x}p(x|z)\overset{\Delta}{=}\hat{x}_{MAP}(z) \tag{3.76}$$

and it is the **maximum of the posteriori density** $p(x|z)$.

To obtain the estimates, it is necessary to determine the a posteriori density $p(x|z)$, but what is usually available is the $p(x|z)$. The a posteriori density can be written in terms of $p(x|z)$ as

$$p(x|z)=\frac{p(z|x)p(x)}{p(z)}=0 \tag{3.77}$$

For each z the $g(z)$ is nothing more than a scale factor, so maximizing $p(x|z)$ is equivalent to **maximizing the numerator** $p(z|x)p(x)$

A necessary condition for the maximum a posteriori estimate is that it satisfy

$$\left.\frac{\partial\left(p(z|x)p(x)\right)}{\partial x}\right|_{x=\hat{x}(z)}=0 \tag{3.78}$$

In many problems where the conditional and marginal densities are exponential in form, a more expedient necessary condition can be obtained by taking the ln of the product of densites before the partial derivative is performed. This will not change the location of the maximum value, and thus it leads to alternative necessary condition for the MAP estimate of X in terms of Z:

$$\left.\frac{\partial(\ln p(z|x))}{\partial x}\right|_{x=\hat{x}_{MAP}(z)} + \left.\frac{\partial(\ln p(x))}{\partial x}\right|_{x=\hat{x}_{MAP}(z)} = 0 \tag{3.79}$$

3.5.2 Examples of Bayes Estimators

Example 3.12

A and X are two random variables. The conditional probability density function for X and the probability density function for A are given as

$$f_X(x \mid a) = ae^{-ax}\mu(x) \qquad and \qquad f_A(a) = \frac{1}{2}[\mu(a-1) - \mu(a-3)]$$

(a) Find the maximum a posteriori estimator for A.

(b) Compute $E\left[\hat{A}_{MAP}(X)\right]$, and say whether the estimator is biased or unbiased.

(c) Compute the variance for your estimator.

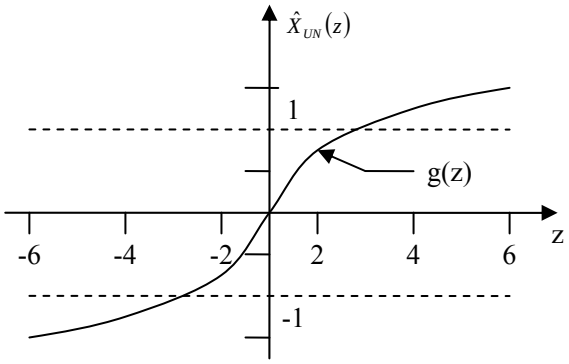

Figure 3.5 Graphical form for Bayes estimator for Example 3-10 with $\sigma_X^2 = \sigma_{N^2} = 1$.

Solution

(a) To obtain the estimator, it is necessary to find the posteriori density $f_A(a \mid x)$, which can be determined as

$$f(a \mid x) = \frac{f_X(x \mid a) f_A(a)}{f_X(x)} = \frac{ae^{-ax}\mu(x)[\mu(a-1) - \mu(a-3)]}{f_X(x)}$$

Since the $f_X(x)$ is just a scale factor and does not depend on a, we need to select the a that maximizes the numerator with respect to a. Figure 3.6 shows for three separate regions for x the plot of $f(a \mid x)$ as a function of a. The maximum is located at different values of a depending on whether the maximum of the ae^{-xa}, which occurs at $1/x$, is in the interval $[1,3]$ or on the boundary.

From Figure 3.6 it is seen that if $1/x<1$, the maximum occurs at the boundary, $a=1$; if $1<1/x<3$, the maximum occurs at $a=1/x$; and if $1/x>3$, then the maximum occurs at the boundary, $a=3$. Thus we can write the maximum a posteriori estimates as follows:

$$\hat{A}_{MAP}(x)=\begin{cases}1, & 0<1/x<1\\ 1/x, & 1<1/x<3\\ 3, & 1/x>3\end{cases}$$

Since $x>0$, the inequalities can be reversed and rearranging gives the $\hat{A}_{MAP}(x)$ as

$$\hat{A}_{MAP}(x)=\begin{cases}3, & 0<x<1/3\\ 1/x, & 1/3<x<1\\ 1, & x>1\end{cases}$$

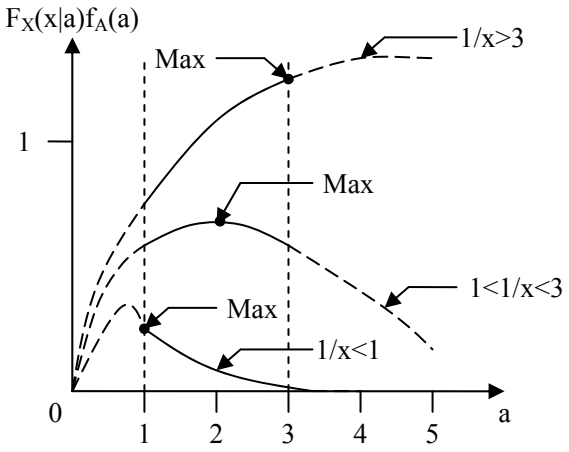

Figure 3.6 Plot of $f_X(x\,|\,a)f_A(a)$ versus a for Example 3-11.

(b) The expected value of the estimator can be written as

$$E\left[\hat{A}_{MAP}(X)\right]=\int_{-\infty}^{\infty}\hat{A}_{MAP}(x)f_X(x)dx$$
$$=\int_0^{1/3}3f_X(x)dx+\int_{1/3}^{1}(1/x)f_X(x)dx+\int_1^{\infty}1\,f_X(x)dx$$

where the $f_X(x)$ is obtained by integrating the product of the conditional density by the marginal as

$$f_X(x)=\int_{-\infty}^{\infty}f_X(x\,|\,a)f_A(a)da$$
$$=\int_1^3 ae^{-ax}\mu(x)\frac{1}{2}da=\frac{-e^{-ax}}{2x}\left(a+\frac{1}{x}\right)\Bigg]_1^3$$
$$=\frac{1}{2}\left(-\frac{3}{x}e^{-3x}-\frac{1}{x^2}e^{-3x}+\frac{1}{x}e^{-x}-\frac{1}{x^2}e^{-x}\right)\mu(x)$$

Using this $f_X(x)$ in the expected value above, and some creative integration.

We obtain the expected value of our estimator as $E\left[\hat{A}_{MAP}(X)\right]=1.635584.$

Since $E[A]$ is easily seen to be 2, this estimator is biased.

(c) The variance of our estimator is easier written down than evaluated as

$$\begin{aligned}\sigma^2_{\hat{A}_{MAP}(X)} &= E\left[\left(A-\hat{A}_{MAP}(X)\right)^2\right] \\ &= \int_1^3\int_0^{1/3}(a-3)^2 f_X(x\mid a)f_A(a)dxda \\ &+\int_1^3\int_{1/3}^{1}(a-1/x)^2 f_X(x\mid a)f_A(a)dxda \\ &+\int_1^3\int_1^{\infty}(a-1)^2 f_X(x\mid a)f_A(a)dxda\end{aligned}$$

Where $f_X(x\mid a)$ and $f_A(a)$ are as given in the problem assignment.

3.6 Estimation of Nonrandom Parameters

Suppose we are trying o estimate the parameter a as a function, $g(z)$, of some observation Z which we can express as $\hat{a}(Z)=g(Z)$. The Z is random and somehow related to a. The $\hat{a}(Z)$ is a random variable because it is a function of a random variable Z. An example of its probability density function is shown in Figure3.7. Intuitively it seems that for $\hat{a}(Z)$ to be a good estimate of the parameter a, the density for $\hat{a}(Z)$ should have a small spread or variance and be centered around the variable being estimated. Such a property can be expressed in terms of the $E[\hat{a}(Z)]$ and the variance $\sigma_{\hat{a}}{}^2 = E[\hat{a}(Z)-E[\hat{a}(Z)]]^2$.

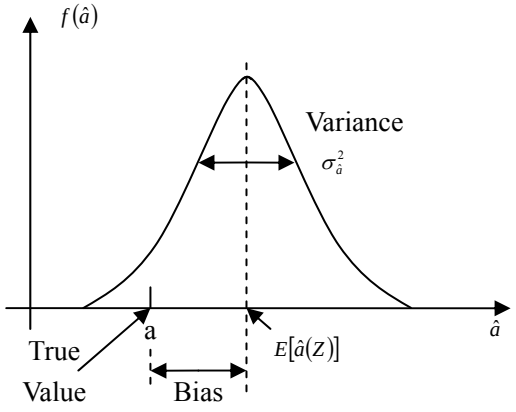

Figure 3.7 Probability density function for $\hat{a}(Z)$.

An estimator of a **nonrandom variable** a is said to be an **unbiased estimator** of a if $E[\hat{a}(Z)]=a$. Therefore we usually seek estimates that are unbiased and have the smallest possible variance. Say that we are estimating a nonrandom parameter a from a vector of random observations Z

and that the variance of our estimator is defined by $E[(a-\hat{a}(Z))^2]$. For even simple problems this variance may be complicated to obtain, and so bounds on estimates are important. The **Cramer-Rao bound** is the most important lower bound on the variance. It is given by

$$E[(a-\hat{a}(Z))^2] \geq \frac{1}{E^2[\partial \ln p(z;a)/\partial a]} \tag{3.80}$$

Any estimator whose variance satisfies the lower bound given in (3.80) is called an **efficient estimator** of a.

3.6.1 Maximum Likelihood Estimation

A random variable Z is observed that is a function of some unknown but nonrandom parameter a. For Example, a constant unknown variable a is observed in additive fashion as $Z = a + X$, where the random variable X has a known statistical characterization and an estimate of a is desired as a function of Z. The maximum likelihood estimation procedure suggests one way to solve for an estimate but not in the sense of an optimality.

The **maximum likelihood estimator** $\hat{a}_{ml}(Z)$ of a nonrandom parameter a in terms of an observed random variable Z is defined to be the a that maximizes the **likelihood function** $f(z;a)$, which is defined as the probability density function for the observation Z in terms of a nonrandom parameter a that could be a vector. Mathematically the function form for the estimator is expressed in terms of $f(z;a)$, which is viewed as a function of a and can be written as

$$\hat{a}_{ml} = \max_a f(z;a) \tag{3.81}$$

Extending this concept to multiple observations, the maximum likelihood estimate of a nonrandom parameter in terms of observation Z_1, Z_2,…, Z_n is defined as the a that maximizes the likelihood function $f(z_1, z_2,..., z_n; a)$ and given by

$$\hat{a}_{ml}(z_1, z_2,..., z_n) = \max_a f(z_1, z_2,..., z_n; a) \tag{3.82}$$

Furthermore, if we have multiple observations and are estimating a parameter vector, $\mathbf{a} = [a_1, a_2,..., a_M]$, rather than a single variable or parameter, the maximum likelihood estimate of $\mathbf{a}$ is given by

$$\hat{\mathbf{a}}_{ml}(z_1, z_2,..., z_n) = \max_a f(z_1, z_2,..., z_n; a) \tag{3.83}$$

3.6.2 Maximum Likelihood Estimation Examples

Example 3.13

The probability density function for an observation Z_i in terms of a nonrandom parameter a is given by

$$f(z_i;a)=ae^{-az_i}\mu(z_i)$$

Say there are n independent observations Z_1, Z_2,…, Z_n.

(a) Find the maximum likelihood estimate $\hat{a}_{ml}$ of a.

(b) Determine the $E[\hat{a}_{ml}]$.

Solution

(a) The likelihood function is a product of the likelihood for each measurement because of the independence of observations. It is determined as follows:

$$\begin{aligned} f(z_1,z_2,...,z_n;a) &= \prod_{i=1}^{n} f(z_i;a) \\ &= a^n \exp\left(-a\sum_{i=1}^{n} z_i\right)\mu(z_1)\mu(z_2)\cdots\mu(z_n) \end{aligned}$$

A necessary condition for the maximum of the likelihood function is that its partial derivative with respective to a be set equal to zero, which is

$$\begin{aligned} \frac{\partial f(z_1,z_2,...,z_n;a)}{\partial a} &= \frac{\partial}{\partial a}\left(a^n \exp\left(-a\sum_{i=1}^{n} z_i\right)\mu(z_1)\mu(z_2)\cdots\mu(z_n)\right) \\ &= n(a^{n-1})\exp\left(-a\sum_{i=1}^{n} z_i\right)-\left(a^n \exp\left(-a\sum_{i=1}^{n} z_i\right)\right)\sum_{i=1}^{n} z_i \\ &= a^{n-1}\exp\left(-a\sum_{i=1}^{n} z_i\right)\left[n-a\sum_{i=1}^{n} z_i\right]=0 \end{aligned}$$

Thus we have relative extremum at $a=0$ due to the a^{n-1} term and at $a=n\Big/\sum_{i=1}^{n} Z_i$ because of the term in square brackets. But $a=0$ is a minimum rather than a maximum, and a maximum occurs for the bracketed term. Therefore the maximum likelihood estimator becomes

$$\hat{a}_{ML}(Z_1,Z_2,...,Z_n)=n\Big/\sum_{i=1}^{n} Z_i$$

(b) The expected value of $\hat{a}_{ML}(Z_1,Z_2,...,Z_n)$ is given by:

$$E[\hat{a}_{ML}(Z_1,Z_2,...,Z_n)]=E\left[n\Big/\sum_{i=1}^{n} Z_i\right]=nE\left[1\Big/\sum_{i=1}^{n} Z_i\right]$$

The expected value of the term in brackets is not the ratio of expected values. Its calculation is facilitated by defining a new random variable Z as

$$Z=1\Big/\sum_{i=1}^{n} Z_i$$

The general approach to determining the required expected value is first to find $f_Z(z)$ by using the Laplace transform (since the density will be a convolution), then to use the transformation theorem to

obtain $f_W(w)$, where $W = 1/Z$, and finally to compute the expected value of W. It is left to the reader to show that

$$E[\hat{a}_{ML}(Z_1, Z_2, ..., Z_n)] = a$$

The estimator is unbiased because its expected value is the unknown parameter or variable being estimated.

Example 3.14

Let a_1 and a_2 be unknown nonrandom parameters observed in terms of

$$Z_1 = a_1 + a_2 + N_1$$
$$Z_2 = a_1 - a_2 + N_2$$

Where N_1 and N_2 are independent Gaussian random variables with zero mean and equal variances of σ^2. Find the maximum likelihood estimates of a_1 and a_2 in terms of Z_1 and Z_2, and determine the expected value of your estimates.

Solution

The likelihood function $f(z_1, z_2; a_1, a_2)$ is given as

$$f(z_1, z_2; a_1, a_2) = \frac{1}{2\pi\sigma^2} \exp\left\{ \frac{-\frac{1}{2}[z_1 - (a_1 + a_2)]^2}{\sigma^2} - \frac{\frac{1}{2}[z_2 - (a_1 - a_2)]^2}{\sigma^2} \right\}$$

Since the $f(z_1, z_2; a_1, a_2)$ is exponential, it is convenient to first take the ln before performing the maximization. The location of the maximum will not be changed if any monotonic function is used. This leads to two necessary conditions for the extremum:

$$\frac{\partial}{\partial a_1} \ln f(z_1, z_2; a_1, a_2) = \frac{\partial}{\partial a_1} \left\{ \frac{-\frac{1}{2}[z_1 - (a_1 + a_2)]^2}{\sigma^2} - \frac{\frac{1}{2}[z_2 - (a_1 - a_2)]^2}{\sigma^2} \right\}$$
$$= [z_1 - (a_1 + a_2)] + [z_2 - (a_1 - a_2)] = 0$$

$$\frac{\partial}{\partial a_2} \ln f(z_1, z_2; a_1, a_2) = \frac{\partial}{\partial a_2} \left\{ \frac{-\frac{1}{2}[z_1 - (a_1 + a_2)]^2}{\sigma^2} - \frac{\frac{1}{2}[z_2 - (a_1 - a_2)]^2}{\sigma^2} \right\}$$
$$= [z_1 - (a_1 + a_2)] - [z_2 - (a_1 - a_2)] = 0$$

Simplifying the equations above gives simultaneous equations as

$$z_1 + z_2 - 2a_1 = 0$$
$$z_1 - z_2 - 2a_2 = 0$$

These equations are easily solved to get

$$\hat{a}_1(z_1,z_2)=\frac{1}{2}(z_1+z_2),\qquad \hat{a}_2(z_1,z_2)=\frac{1}{2}(z_1-z_2)$$

Since $E[N_1]=E[N_2]=0$, the expected values of our estimates can be determined as

$$\begin{aligned}E[\hat{a}_1(Z_1,Z_2)]&=E\left[\frac{1}{2}(Z_1+Z_2)\right]\\&=E[(a_1+N_1+N_2)]=a_1\\E[\hat{a}_2(Z_1,Z_2)]&=E\left[\frac{1}{2}(Z_1-Z_2)\right]\\&=E[(a_2+N_1-N_2)]=a_2\end{aligned}$$

Thus we have shown that the estimates are unbiased for this problem, since there expected values equal the parameters to be estimated. The variances of our estimates for this problem are easily found as

$$\begin{aligned}\sigma^2_{\hat{a}_1}&=E[(\hat{a}_1(Z_1,Z_2)-a_1)^2]=E\left[\left(\frac{1}{2}(Z_1+Z_2)-a_1\right)^2\right]\\&=E\left[\frac{1}{2}(N_1+N_2)^2\right]=\sigma^2\\\sigma^2_{\hat{a}_2}&=E[(\hat{a}_2(Z_1,Z_2)-a_2)^2]=E\left[\left(\frac{1}{2}(Z_1-Z_2)-a_2\right)^2\right]\\&=E\left[\frac{1}{2}(N_1-N_2)^2\right]=\sigma^2\end{aligned}$$

3.7 Summary

In this chapter basic procedures were presented for optimally estimating random variables from other random variables where optimality was with respect to a given performance measure. Emphasis was placed on the minimum mean squared error performance measure, not only because of its tractability but because many practical problems do not have a full characterization of the involved random variables.

It was shown that for minimum mean squared error estimation of a random variable by a linear combination of other random variables, the only characterizations of the random variables required are the means, auto-correlations, and cross-correlations. The orthogonality principle provided another interpretation and the equations necessary to obtain optimum estimate in the MMSE sense.

For nonlinear MMSE estimation, it was shown that the mean of the conditional density was the needed function, which involved almost a total characterization of the random variables. For the special case of Gaussian random variables, it was shown that the optimal linear estimate is the same as the optimum nonlinear estimate when finding MMSE estimates.

Examples were presented using the Bayes performance measure for special cases of squared error, absolute error, and uniform error cost functions. It was shown that these estimates are correspondingly the

mean, median, and mode of the involved a posteriori probability density, which involves almost a total characterization.

The topic of estimating nonrandom variables was introduced as there are many problems in which the estimate parameters or variables to be estimated are not random. The one method presented was the maximum likelihood procedure, and several examples were given. The concept of variance and bias in describing the performance of estimators of nonrandom parameters or variables was explored, with the result being that we would like our estimates to have a small variance and be unbiased.

----------This is the end of Chapter03----------