From Papoulis

 27

7-4. Mean-square estimation; the orthogonality principle

We shall now be concerned with the question of estimating a r.v. $\mathbf{x}$ by a constant or by a function of another r.v. $\mathbf{y}$. This problem will be reexamined in Sec. 8-2 and in Chap. 11. In the following discussion we consider its meaning and introduce the notion of mean square (abbreviation: m.s.) estimation. The formulation and solution of the problem will be in terms of probabilities (conceptual). However, a brief explanation in terms of repeated trials (physical) might be helpful.

Frequency interpretation. A r.v. $\mathbf{x}$ is defined on a certain experiment. Its distribution $F(x)$ is given, and so is its value $\mathbf{x}(\zeta)$ for each outcome ζ. This does not, of course, mean that if the corresponding physical experiment is performed, one will know in advance the resulting value $\mathbf{x}(\zeta)$ of $\mathbf{x}$. The outcome ζ of a particular trial might be any element of $\mathcal{S}$. The question arises whether, guided by $F(x)$, we could "guess" a value a for $\mathbf{x}(\zeta)$. This is the problem of estimating the r.v. $\mathbf{x}$ by a constant. Suppose that a is somehow selected. At each trial we commit an error,

$$\mathbf{x}(\zeta) - a \tag{7-89}$$

and our problem is to find the particular a that will make this error "small." If by "small" we mean that the average of $\mathbf{x}(\zeta) - a$ in a long run of trials should be close to zero,

$$\frac{\mathbf{x}(\zeta_1) - a + \cdots + \mathbf{x}(\zeta_n) - a}{n} \simeq 0$$

then [see (5-22)] a should equal the expected value of $\mathbf{x}$.

However, depending on the nature of the problem, one might prefer some other criterion for selecting a, for example, the minimization of the average of $|\mathbf{x}(\zeta) - a|$. In this case, a should equal the median of $\mathbf{x}$ (see Prob. 5-3). In our analysis we shall deal only with m.s. estimations. This means that a should be so selected that the average of

$$[\mathbf{x}(\zeta) - a]^2$$

is minimum. This criterion is, in general, useful, but it is primarily chosen because it leads to simple results. We shall soon see that the best a is again the expected value of $\mathbf{x}$,

$$a = E\{\mathbf{x}\} \tag{7-90}$$

The estimation of $\mathbf{x}$ can be improved if one has access to the values of another r.v. $\mathbf{y}$. We elaborate: It is assumed that at each trial we "observe" the resulting value $\mathbf{y}(\zeta)$ of $\mathbf{y}$ and want $\mathbf{x}(\zeta)$ estimated on the basis of this observation. If $\mathbf{x}$ and $\mathbf{y}$ are independent, then knowledge of $\mathbf{y}(\zeta)$ is of no help in the estimate of $\mathbf{x}$. In this case $\mathbf{x}$ is again estimated by a constant. However, if $\mathbf{x}$ and $\mathbf{y}$ are not independent, then it might be best to use for an estimate of $\mathbf{x}$ not the same number at each trial, but a quantity that depends on the observed $\mathbf{y}(\zeta)$. In other words, we want $\mathbf{x}$ estimated by a function $g(\mathbf{y})$, and our problem is to find the best $g(\mathbf{y})$.

One might argue that if $\mathbf{y}(\zeta)$ is observed, then the particular outcome ζ is known; hence $\mathbf{x}(\zeta)$ can be *predicted* exactly. This is not so. The same number $y = \mathbf{y}(\zeta)$ might result from several outcomes ζ,

$$y = \mathbf{y}(\zeta_1) = \cdots = \mathbf{y}(\zeta_n) = \cdots \tag{7-91}$$

and for each such ζ the corresponding values of $\mathbf{x}$ might be different. Hence, having observed $\mathbf{y}(\zeta)$ at a given trial, we cannot, in general, predict $\mathbf{x}(\zeta)$, but only estimate† it.

The foregoing reasoning leads to the following conclusion: The fact that $\mathbf{y}(\zeta) = y$ is specified means that the outcome ζ of our trial is not any element of $\mathcal{S}$, but only one of the elements ζ_i in (7-91). In other words, we are asking for an estimate of $\mathbf{x}$ in the subset $\{\mathbf{y} = y\}$ of our space. In this set, $\mathbf{y}(\zeta)$ is a constant, and our problem is to estimate $\mathbf{x}$ by the *constant* $g[\mathbf{y}(\zeta)]$. Changing probabilities into conditional probabilities, we conclude, as in (7-90), that the best m.s. estimate of $\mathbf{x}$ is its expected value

$$g(\mathbf{y}) = E\{\mathbf{x}|\mathbf{y}\}$$

We shall soon see that the above loose conclusions can be strictly established in the conceptual world of probabilities.

Mean-square estimation of a random variable by a constant. We start with the following simple but basic problem: Find a constant a such that

$$E\{(\mathbf{x} - a)^2\} = \int_{-\infty}^{\infty} (x - a)^2 f(x)\, dx$$

is minimum. We maintain that

$$a = E\{\mathbf{x}\} = \eta_x = \int_{-\infty}^{\infty} x f(x)\, dx \tag{7-92}$$

Indeed, expanding, we have

$$E\{(\mathbf{x} - a)^2\} = a^2 - 2aE\{\mathbf{x}\} + E\{\mathbf{x}^2\}$$

The derivative with respect to a equals zero for $a = E\{\mathbf{x}\}$, and (7-92) follows. Thus the constant η_x has the properties‡ that: The expected value of $\mathbf{x} - \eta_x$ equals zero; the expected value of $(\mathbf{x} - \eta_x)^2$ is minimum.

Nonlinear mean-square estimation of y in terms of x. We now want to estimate the r.v. $\mathbf{y}$ by a suitable function $g(\mathbf{x})$ of $\mathbf{x}$ so that the m.s. estimation error

$$E\{[\mathbf{y} - g(\mathbf{x})]^2\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [y - g(x)]^2 f(x,y)\, dx\, dy \tag{7-93}$$

is minimum (we reversed the role of $\mathbf{x}$ and $\mathbf{y}$).

† In the literature the expression "prediction of $\mathbf{x}$ in terms of $\mathbf{y}$" is often used; from the above we see that "estimation" is a more appropriate term.

‡ The above result corresponds to the well-known fact that the moment of inertia with respect to the center of gravity is smaller than with respect to any other point.

Theorem. The function $g(\mathbf{x})$ that minimizes (7-93) is the conditional expected value of $\mathbf{y}$, assuming $\mathbf{x}$:

$$g(\mathbf{x}) = E\{\mathbf{y}|\mathbf{x}\} \tag{7-94}$$

Proof. Since

$$f(x,y) = f(y|x)f(x)$$

we have

$$E\{[\mathbf{y} - g(\mathbf{x})]^2\} = \int_{-\infty}^{\infty} f(x) \int_{-\infty}^{\infty} [y - g(x)]^2 f(y|x)\, dy\, dx$$

The integrand above is nonnegative; therefore, in order to minimize the double integral, it suffices to minimize

$$\int_{-\infty}^{\infty} [y - g(x)]^2 f(y|x)\, dy$$

for every x. For a given x, this integral is the second moment of the conditional density $f(y|x)$ with respect to the *constant* $g(x)$. As we know from (7-92), this moment is minimum if

$$g(x) = \int_{-\infty}^{\infty} y f(y|x)\, dy = E\{\mathbf{y}|x\}$$

and (7-94) follows.

Thus (7-94) is a simple extension of (7-92) in the probability space conditioned by $\{\mathbf{x} = x\}$. This conclusion can also be drawn, after some thought, from [see (7-59)]

$$E\{[\mathbf{y} - g(\mathbf{x})]^2\} = E\{E\{[\mathbf{y} - g(\mathbf{x})]^2|\mathbf{x}\}\}$$

The function

$$g(x) = E\{\mathbf{y}|x\}$$

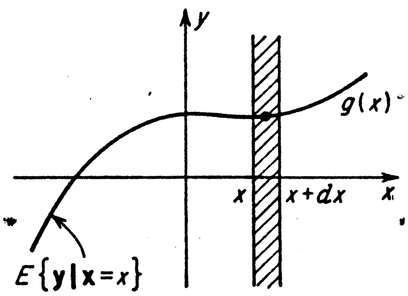

Fig. 7-18

is known as *regression curve* (Fig. 7-18). It is the locus of the centers of gravity of the masses on the strips $(x, x + dx)$. If these masses are near $g(x)$, then the m.s. error

$$E\{[\mathbf{y} - E\{\mathbf{y}|\mathbf{x}\}]^2\} \tag{7-95}$$

is small.

Independent Random Variables. If $\mathbf{x}$ and $\mathbf{y}$ are independent, then (see page 182)

$$E\{\mathbf{y}|\mathbf{x}\} = E\{\mathbf{y}\}$$

Hence the best m.s. estimate of $\mathbf{y}$ in terms of $\mathbf{x}$ is $E\{\mathbf{y}\}$. Thus knowledge of $\mathbf{x}$ does not help in the estimation of $\mathbf{y}$.

Linear Mean-square Estimation; the Orthogonality Principle

The solution $E\{\mathbf{y}|\mathbf{x}\}$ of the nonlinear estimation problem looks simple enough. However, the actual evaluation of $E\{\mathbf{y}|\mathbf{x}\}$ is not simple at all. One must determine this function for every $\mathbf{x}$. The difficulties become severe if more than one r.v. are involved (Chap. 11). A much easier problem is the estimation of $\mathbf{y}$ by a linear function

$$a\mathbf{x} + b$$

of $\mathbf{x}$. We now seek not a function, but merely the two constants a and b that minimize

$$E\{[\mathbf{y} - (a\mathbf{x} + b)]^2\}$$

The resulting error is of course larger than the corresponding error in the nonlinear estimation; however, this is often compensated by the simplicity of the solution.

Theorem. The constants a and b that minimize the m.s. error

$$e = E\{[\mathbf{y} - (a\mathbf{x} + b)]^2\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y - ax - b)^2 f(x,y)\, dx\, dy \qquad (7\text{-}96)$$

are given by

$$a = \frac{r\sigma_y}{\sigma_x} \qquad b = E\{\mathbf{y}\} - aE\{\mathbf{x}\} \qquad (7\text{-}97)$$

and the resulting minimum error e_m by

$$e_m = \sigma_y^2(1 - r^2) \qquad (7\text{-}98)$$

where r is the correlation coefficient of $\mathbf{x}$ and $\mathbf{y}$ [see (7-66)].

Proof. Suppose that a is specified. The value of b that minimizes e is the best m.s. estimate of the r.v. $\mathbf{y} - a\mathbf{x}$ by a constant; hence [see (7-92)]

$$b = E\{\mathbf{y} - a\mathbf{x}\} = \eta_y - a\eta_x$$

With b so determined, we now have

$$E\{(\mathbf{y} - a\mathbf{x} - b)^2\} = E\{[(\mathbf{y} - \eta_y) - a(\mathbf{x} - \eta_x)]^2\} = \sigma_y^2 - 2r\sigma_x\sigma_y a + \sigma_x^2 a^2$$

The last quantity is minimum for

$$a = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = \frac{r\sigma_y}{\sigma_x}$$

and (7-97) follows. Inserting the value of a in the above quadratic, we find

$$e_m = \sigma_y^2 - 2r^2\sigma_y^2 + r^2\sigma_y^2 = \sigma_y^2(1 - r^2)$$

$$r = \frac{E\{(x-\eta_x)(y-\eta_y)\}}{\sqrt{E\{(x-\eta_x)^2\}\,E\{(y-\eta_y)^2\}}} = \frac{\mu_{11}}{\sigma_x\sigma_y}$$

30

The above result will now be stated in a more basic form. We shall assume for simplicity that

$$E\{\mathbf{x}\} = E\{\mathbf{y}\} = 0$$

Orthogonality principle. The constant a that minimizes the m.s. error

$$e = E\{(\mathbf{y} - a\mathbf{x})^2\}$$

is such that $\mathbf{y} - a\mathbf{x}$ is orthogonal to $\mathbf{x}$; that is,

$$E\{(\mathbf{y} - a\mathbf{x})\mathbf{x}\} = 0 \tag{7-99}$$

and the minimum m.s. error is given by

$$e_m = E\{(\mathbf{y} - a\mathbf{x})\mathbf{y}\} \tag{7-100}$$

Proof. The above can be deduced from (7-97) and (7-98); it will be instructive, however, to give a second proof. This proof can be simply extended to complex r.v. (see Sec. 11-2), whereas the differentiation presents certain complications.

Suppose that a is such that $E\{(\mathbf{y} - a\mathbf{x})\mathbf{x}\} = 0$. We maintain that the resulting error e is minimum. Indeed, for any A, we have

$$\begin{aligned}E\{(\mathbf{y} - A\mathbf{x})^2\} &= E\{[(\mathbf{y} - a\mathbf{x}) + (a - A)\mathbf{x}]^2\}\\ &= E\{(\mathbf{y} - a\mathbf{x})^2\} + 2(a - A)E\{(\mathbf{y} - a\mathbf{x})\mathbf{x}\} + (a - A)^2E\{\mathbf{x}^2\}\end{aligned}$$

But the second term in the last expression is zero, and the last nonnegative; hence

$$E\{(\mathbf{y} - A\mathbf{x})^2\} \geq E\{(\mathbf{y} - a\mathbf{x})^2\}$$

and our statement is proved. The minimum error is given by

$$e_m = E\{(\mathbf{y} - a\mathbf{x})^2\} = E\{(\mathbf{y} - a\mathbf{x})\mathbf{y}\} - aE\{(\mathbf{y} - a\mathbf{x})\mathbf{x}\}$$

and (7-100) follows because the last term is zero.

From (7-99) we conclude that

$$a = \frac{E\{\mathbf{xy}\}}{E\{\mathbf{x}^2\}} \tag{7-101}$$

Inserting into (7-100), we obtain

$$e_m = E\{\mathbf{y}^2\} - aE\{\mathbf{xy}\} = E\{\mathbf{y}^2\} - \frac{E^2\{\mathbf{xy}\}}{E\{\mathbf{x}^2\}} \tag{7-102}$$

The m.s. error can also be written in the form

$$e_m = E\{\mathbf{y}^2\} - E\{(a\mathbf{x})^2\} \tag{7-103}$$

We remark that

$$e_m \geq E\{[\mathbf{y} - E\{\mathbf{y}|\mathbf{x}\}]^2\}$$

31

We shall now extend briefly the results of Sec. 7-4 to several r.v. This discussion will be resumed in Chap. 11. We are given the $n+1$ r.v.

$$\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n \tag{8-29}$$

and we want to estimate $\mathbf{x}_0$ by a function $g(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of the other r.v. so as to minimize the m.s. error:

$$E\{[\mathbf{x}_0 - g(\mathbf{x}_1, \ldots, \mathbf{x}_n)]^2\} \tag{8-30}$$

Reasoning as in (7-94), we can easily show that

$$g(\mathbf{x}_1, \ldots, \mathbf{x}_n) = E\{\mathbf{x}_0 | \mathbf{x}_1, \ldots, \mathbf{x}_n\} \tag{8-31}$$

The above expected value is given by (8-17). Thus, to solve the nonlinear m.s. estimation problem, we need to know the joint density of the r.v. $\mathbf{x}_0, \ldots, \mathbf{x}_n$.

Linear mean-square estimation. The estimation problem is considerably simplified if one seeks an estimate of $\mathbf{x}_0$ by a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_n$. In this case the problem is to find n constants $a_1, \ldots, a_n$ such that the m.s. error

$$e = E\{[\mathbf{x}_0 - (a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n)]^2\} \tag{8-32}$$

is minimum. It turns out that these constants can be determined in terms of the second moments

$$R_{ij} = E\{\mathbf{x}_i\mathbf{x}_j\}$$

of the given r.v. If $E\{\mathbf{x}_i\} = 0$, then R_{ij} is the covariance of the r.v. $\mathbf{x}_i$ and $\mathbf{x}_j$.

Orthogonality Principle. The constants a_i that minimize e are such that the error

$$\mathbf{x}_0 - (a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n)$$

is orthogonal to $\mathbf{x}_1, \ldots, \mathbf{x}_n$; that is,

$$\boxed{E\{[\mathbf{x}_0 - (a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n)]\mathbf{x}_i\} = 0 \qquad i = 1, \ldots, n} \tag{8-33}$$

Proof. The m.s. error e is a function of $a_1, \ldots, a_n$, and to minimize it we differentiate with respect to a_i:

$$\frac{\partial e(a_1, \ldots, a_n)}{\partial a_i} = \frac{\partial E\{[\mathbf{x}_0 - (a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n)]^2\}}{\partial a_i} = 0 \qquad i = 1, \ldots, n$$

Writing the above expected value as an integral of the form (8-15), we see that the order of differentiation and expected value can be interchanged; the result is

$$\frac{\partial e}{\partial a_i} = -2E\{[\mathbf{x}_0 - (a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n)]\mathbf{x}_i\} = 0$$

and (8-33) follows.

It is easy to see by expanding the square in (8-32) and using (8-33) that the minimum m.s. error is given by

$$e_m = E\{[\mathbf{x}_0 - (a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n)]\mathbf{x}_0\} = R_{00} - (a_1R_{01} + \cdots + a_nR_{0n}) \tag{8-34}$$