CARPATHIAN J. MATH.
20 (2004), No. 1, 31 - 38

The orthogonality principle and conditional densities

CRISTINA IOANA FĂTU

ABSTRACT. Let $X, Y \in L^2(\Omega, K, P)$ be a pair of random variables, where $L^2(\Omega, K, P)$ is the space of random variables with finite second moments. If we suppose that X is an observable random variable but Y is not, than we wish to estimate the unobservable component Y from the knowledge of observations of X. In this paper, using some definitions and properties of the estimators we shall present some results relative to the mean-square estimation.

1. CONVERGENCE IN THE MEAN-SQUARE

Let (Ω, K, P) be a probability space and $X, X_1, X_2, ...$ a sequence of random variables defined on this space. There are a number of ways in which the sequence might converge as $n \to \infty$. In the next we will recall some from them [5],[6].

Definition 1.1. The sequence $X_1, X_2, ...$ of random variables converges in probability to the random variables X if for every $\varepsilon > 0$, we have

$$(1.1) \qquad \lim_{n\to\infty} P\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\} = 0 \quad \text{or}$$
$$P\{\omega : |X_n(\omega) - X(\omega)|\} \to 0, n \to \infty.$$

Symbolically written as $X_n \xrightarrow{P} X$ or $p \lim_{n\to\infty} X_n = X$.

Let (Ω, K, P) be a probability space and $\mathcal{F}(\Omega, K, P)$ the family of all random variables defined on (Ω, K, P). Let

$$(1.2) \qquad L^p = L^p(\Omega, K, P) = \{X \in \mathcal{F}(\Omega, K, P) \mid E(|X|^p) < \infty\}, p \in \mathbb{N}^*$$

be the set of random variables with finite moments of order p, that is,

$$(1.2a) \quad \beta_p = E(|X|^p) = \int_{\mathbb{R}} |x|^p \, dF(x) < \infty, p \in \mathbb{N}^*,$$

where

$$(1.3) \qquad F(x) = P(X < x), x \in \mathbb{R}$$

is the distribution function of the random variable X.

Remark 1.1. The set $L^p(\Omega, K, P)$ represents a linear space.

Received: 26.09.2004; In revised form: 12.01.2005
2000 *Mathematics Subject Classification.* 62H10, 62H12.
Key words and phrases. *Estimation, mean-square estimation, conditional means, orthogonality principle, conditional densities.*

Indeed, if $X_1,\ X_2 \in L^p(\Omega, K, P)$ and $c_1, c_2 \in \mathbb{R}$, then and the random variable X, defined by the relation

$$X = c_1X_1 + c_2X_2,\ \forall c_1, c_2 \in \mathbb{R}, \tag{1.4}$$

is also from the set $L^p(\Omega, K, P)$, if we have in view the Minkowski's inequality

$$[E(|X_1 + X_2|^p)]^{\frac{1}{p}} \leq [E(|X_1|^p)]^{\frac{1}{p}} + [E(|X_2|^p)]^{\frac{1}{p}}, p \geq 1. \tag{1.5}$$

Among the spaces $L^p = L^p(\Omega, K, P)$, $p \geq 1$, an important role is played by the space $L^2 = L^2(\Omega, K, P)-$ *the space of random variables with finite second moments.*

Definition 1.2. [5] If $X, Y \in L^2(\Omega, K, P)$, then the distance in mean square between X and Y, denoted by $d_2(X, Y)$, is defined by the equality

$$d_2(X, Y) = \|X - Y\| = [E(|X - Y|^2)]^{1/2}. \tag{1.6}$$

Remark 1.2. It is easy to verify that $d_2(X, Y)$ satisfies the following conditions:

$$\begin{cases} 1^0 & d_2(X,Y) = \|X - Y\| \geq 0, \forall X, Y \in L^2(\Omega, K, P); \\ 2^0 & d_2(X,X) = \|X - X\| = 0, \forall X \in L^2(\Omega, K, P); \\ 3^0 & d_2(X,Y) = \|X-Y\| = \|Y-X\| = d_2(Y,X), \forall X, Y \in L^2(\Omega, K, P); \\ 4^0 & d_2(X,Z) \leq d_2(X,Y) + d_2(Y,Z), \forall X, Y, Z \in L^2(\Omega, K, P), \end{cases} \tag{1.7}$$

that is, $d_2(X, Y)$ represents a semi-metric on the linear space L^2.

Definition 1.3. [1], [5] If $(X, X_n, n \geq 1) \subset L^2(\Omega, K, P)$, then about the sequence $(X_n)_{n \in \mathbb{N}^*}$ is said to converge to X in mean square (converge in L ^{2}) if

$$\lim_{n\to\infty} d_2(X_n, X) = \lim_{n\to\infty} E(|X_n - X|^2)^{1/2} = \tag{1.8}$$

$$= \lim_{n\to\infty} E(|X_n - X|^2) = 0. \tag{1.8a}$$

We write

$$l.i.m.X_n = X \ or \ X_n \xrightarrow{m.p.} X, n \to \infty, \tag{1.9}$$

and call X the limit in the mean (or mean square limit) of X_n.

Remark 1.3. [1] If $X \in L^2(\Omega, K, P)$, then

$$Var(X) = E[(X - m)^2] = E[|X - m|^2] = \|X - m\|^2 = d_2^2(X, m), \tag{1.9a}$$

where $m = E(X)$.

2. Mean-square estimation

Consider two random variables X and Y. Suppose that only X can be observed. If X and Y are correlated, we may expect that knowing the value of X allows us to make some inference about the value of the unobserved variable Y. In this case arises an interesting problem, namely that of estimating one random variable with another or one random vector with another.

If we consider *any function* $\widehat{X} = g(X)$ on X, then that is called *an estimator for* Y. A desirable property of any estimator $\widehat{X}$ of Y would be that

$$E(\widehat{X}) = Y, \tag{2.1}$$

i.e., in other words, the average of estimator is the true value. When any estimator satisfies (2.1), it is said to be *unbiased*. *The error* is defined as the difference between the estimator and the true value, that is,

$$e = \widehat{X} - Y. \tag{2.2}$$

If $\widehat{X}$ is *an unbiased* estimator then this error (in the estimate) can be written as

$$e = \widehat{X} - E(\widehat{X}). \tag{2.3}$$

This error is a random variable, since, in general, both $\widehat{X}$ and Y are random in nature. Also, the error may be positive or negative. We cannot minimize the error directly but must choose some arbitrary function of e to minimize. An intuitive and physically pleasing choice is the average mean-square error of the components of e. In other words, we choose to minimize the diagonal terms of the following matrix

$$\mathbf{K}_e = E[(\widehat{X} - Y)(\widehat{X} - Y)^T]. \tag{2.4}$$

In this case $\widehat{X}$ is called *the minimum mean- square error estimator.*

If $\widehat{X}$ is *an unbiased estimator* then the matrix $\mathbf{K}_e$ has the form

$$\mathbf{K}_e = E[(\widehat{X} - E(\widehat{X}))(\widehat{X} - E(\widehat{X}))^T], \tag{2.5}$$

and $\mathbf{K}_e$ is just *the covariance matrix* of the estimator $\widehat{X}$.

In this last case $\widehat{X}$ is called *the minimum variance unbiased estimator.* This type of estimator will be our choice for *the optimum or best estimator.*

Definition 2.1. We say that a function $X^* = g^*(X)$ on X is best estimator in the mean-square sense if

$$E\{[Y - X^*]^2\} = E\{[Y - g^*(X)]^2\} = \inf_g E\{[Y - g(X)]^2\}. \tag{2.6}$$

Theorem 2.1. [1], [3] *Let X, Y be two random variables such that $E(X) = 0$, $E(Y) = 0$ and $\widehat{X}$ a new random variable, $\widehat{X} \in L^2(\Omega, K, P)$, defined as*

$$\widehat{X} = g(X) = a_0 X, \ a_0 \in \mathbb{R}. \tag{2.7}$$

The real constant a_0 that minimize the mean-square error

$$E[(Y - \widehat{X})^2] = E[(Y - a_0 X)^2] \tag{2.8}$$

is such that the random variable $Y - a_0 X$ is orthogonal to X, that is,

$$E[(Y - a_0 X)X] = 0 \tag{2.9}$$

and the minimum mean-square error is given by

$$e_{\min}(Y, \widehat{X}) = e_{\min} = E[(Y - a_0 X)Y], \tag{2.10}$$

where

$$a_0 = \frac{E(XY)}{E(X^2)} = \frac{cov(X,Y)}{\sigma_1^2}. \tag{2.11}$$

Remark 2.1. This theorem represents the orthogonality principle in the case of the linear mean-square estimation, that is, then when $\widehat{X} = g(X)$ is a linear function of X of the form (2.7).

In the next we consider $(n+1)$ random variables

(2.12) $\quad Y, X_1, X_2, ..., X_n \in L^2(\Omega, K, P)$

and we want to estimate Y by a nonlinear function on random vector $\mathbf{X} = (X_1, X_2, ..., X_n)^T$ of the form

(2.13) $\quad \widehat{X}_0 = g_0(\mathbf{X}) = g_0(X_1, X_2, ..., X_n)$

so as to minimize the mean-square error

(2.14) $\quad e = e(Y, \widehat{X}_0) = E[(Y - \widehat{X}_0)^2]$,

that is, to have

$$e_{min}(Y, \widehat{X}_0) = E[(Y - \widehat{X}_0)^2] =$$

(2.15) $$= E\left\{[Y - g_0(X_1, X_2, ..., X_n)]^2]\right\}.$$

Theorem 2.2. *The random variable*

(2.16) $\quad \widehat{X}_0 = g_0(X_1, X_2, ..., X_n) = g_0(\mathbf{X}) =$

2.16a $\quad = E[Y \mid (X_1, X_2, ..., X_n)^T] =$

2.16b $\quad = E(Y \mid \mathbf{X})$,

defined by the conditional expectation of Y with respect to random vector X and with the real values of the form

(2.17) $$M[Y \mid \mathbf{X} = \mathbf{x}] = \int_{-\infty}^{\infty} y f(y \mid \mathbf{x}) dy,$$

for any $n-$ dimensional real point x of the form

(2.17a) $\mathbf{x} = (x_1, x_2, .., x_n)^T \in \mathbb{D}_{\mathbf{x}} = \{\mathbf{x} \in \mathbb{R}^n | f(x_1, .., x_n) = f(\mathbf{x}) > 0\}$,

represents an optimal estimator (the best estimator in the mean-square sense) for the random variable Y, that is,

(2.18) $\quad e_{min}(Y, \widehat{X}_0) = \min_X E[(Y - \widehat{X})^2] =$

2.18a $\quad = E\left\{[Y - g_0(\mathbf{X})]^2]\right\} =$

2.18b $\quad = E\left\{[Y - E(Y \mid \mathbf{X})]^2\right\}.$

Proof. First, we will recall the definition and some very important properties of the conditional mean. □

Definition 2.2. [3] The conditional mean of the random variable Y given the random variable $X = x$, denoted by $E(Y \mid X = x)$, is defined by

(2.19) $\quad E(Y \mid X = x) = E(Y \mid x) =$

2.19a $$= \int_{-\infty}^{\infty} y f(Y \mid X = x) dy =$$

2.19b $$= \int_{-\infty}^{\infty} y f(y \mid x) dy,$$

for any $x \in D_x = \{x \in R \mid f(x) > 0\}$.

Theorem 2.3. [5] *Let* $\widehat{X}$ *be a random variable defined as a nonlinear function of* X, *namely*

(2.20) $\widehat{X} = g(X)$

where $g(x)$ *represents the value of this random variable* $g(X)$ *in the point* x, $x \in D_x$. *Then, the minimum value of the mean-square error, namely,*

(2.20a) $e_{\min} = e_{\min}(Y, \widehat{X}) = E\left\{[(Y - E(Y \mid X)]^2\right\}$

is obtained if

(2.21) $g(X) = E(Y \mid X)$,

where $E(Y \mid X)$ *is the random variable defined by the conditional expectation of* Y *with respect to* X.

Lemma 2.1. [5] *Because the quantity* $E(Y \mid X)$ *is a random variable with the real values of the form* (2.19*b*) *it follows that the expected value of this random variable is equal with the expected value of* Y, *that is,*

(2.22) $E[E(Y \mid X)] = E(Y)$.

The Theorem 2.1 is a generalization of the Theorem 2.2.

In the next we will present a new proof which use this last lemma. For this we write the function $g(\mathbf{X})$ as

(2.23) $\widehat{X} = g(\mathbf{X}) = g_0(\mathbf{X}) + b(g)$,

where the difference

(2.23a) $b(g) = |g(\mathbf{X}) - g_0(\mathbf{X})|$

represents the error of the any estimator $g(\mathbf{X})$ relative to the optimal estimator $g_0(\mathbf{X})$.

Then, the mean-square error can be expressed as

$$\begin{aligned} e = e(Y, \widehat{X}) &= E[(Y - \widehat{X})^2] = \\ &= E\{[Y - g_0(\mathbf{X}) - b(g)]^2\} = \\ &= E\{[Y - E(Y \mid \mathbf{X})]^2 - 2[Y - E(Y \mid \mathbf{X})]b(g)\} + [b(g)]^2\} = \\ &= E\{[Y - E(Y \mid \mathbf{X})]^2\} + E\{[b(g)]^2 - 2E\{[Y - E(Y \mid \mathbf{X})]b(g)\} = \\ &= E\{[Y - E(Y \mid \mathbf{X})]^2\} + E\{[b(g)]^2\}, \end{aligned} \tag{2.24}$$

if we have in view that

$$E\{[Y - E(Y \mid \mathbf{X})]b(g)\} = E[Yb(g)] - \underbrace{E\{E[(Yb(g)) \mid \mathbf{X})]\}}_{\text{(see, Lemma 2.1)}} =$$

2.24a $$= E[Y\delta(g)] - E[Yb(g)] = 0.$$

The relation

(2.25) $E\{[Y - E(Y \mid \mathbf{X})]b(g)\} = 0$,

express the fact that the error vector $\varepsilon = Y - M(Y \mid \mathbf{X})$ is orthogonal to the bias $b(g)$.

Then, from (2.24), we obtain

(2.26) $e = e(Y, \widehat{X}) = E\{[Y - E(Y \mid \mathbf{X})]^2\} + \underbrace{E\{[b(g)]^2\}}_{\geq 0} \geq$

2.26a $= E\{[Y - E(Y \mid \mathbf{X})]^2\} = e_{\min}(Y, \widehat{X}_0).$

Also, we observe that, if $b(g) = 0$, then from (2.23) we can obtain the following equality

(2.27) $\widehat{X} = g(\mathbf{X}) = g_0(\mathbf{X}) = \widehat{X}_0,$

which implies that the minimum mean-square error estimator is

(2.27a) $\widehat{X}_0 = g_0(\mathbf{X}) = E(Y \mid \mathbf{X}).$

Theorem 2.4. *Let X and Y be two random vectors,* $\dim X = \dim Y = n \times 1$*. We suppose that only X can be observed and Y is an unobservable random vector. Let $f(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n)$ be the joint probability density function of $2n-$ dimensional random vector (X, Y). If X and Y are dependent random vectors then, the optimal estimator of the unknown random vector Y, then when the random vector X was observed, is a (possibly nonlinear) function of X, of the form*

(2.28) $\widehat{X}_0 = g_0(\mathbf{X}) = E(\mathbf{Y} \mid \mathbf{X}) =$

2.28a $= [E(Y_1 | \mathbf{X}), E(Y_2 | \mathbf{X}), ..., E(Y_n | \mathbf{X})] =$

2.28b $= [g_0^{(1)}(\mathbf{X}), g_0^{(2)}(\mathbf{X}), ..., g_0^{(n)}(\mathbf{X})]$

and the total minimum mean-square error can be expressed as

(2.29) $e_{\min}(\mathbf{Y}, \widehat{X}_0) = \sum_{i=1}^{n} e_{\min}(Y_i, \widehat{X}_0) =$

2.29a $= \sum_{i=1}^{n} E\{[Y_i - g_0^{(i)}(\mathbf{X})]^2\},$

where

(2.29b) $g_0^{(i)}(X) = E(Y_i \mid X),\ i = \overline{1, n}.$

Proof. Because the random vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ has n elements, which are the unidimensional random variables $Y_1, Y_2, ..., Y_n$, it follows that, in the next, we must to find, using the observed values of the $n-$dimensional random vector $\mathbf{X}$, an optimal estimator for each of them.

Thus, in accordance with the Theorem 2.2, for each random variable $Y_i, i = \overline{1, n}$, the optimal estimator $\widehat{X}_0^{(i)}$ has the form

(2.30) $\widehat{X}_0^{(i)} = g_0^{(i)}(\mathbf{X}) = E(Y_i \mid \mathbf{X}),\ i = \overline{1, n},$

and the individual minimum mean-square error can be expressed as

(2.31) $e_{\min}^{(i)}(Y_i, \widehat{X}_0^{(i)}) = E\{[Y_i - E(Y_i \mid \mathbf{X})]^2\} =$

2.31a $= E\{[Y_i - g_0^{(i)}(\mathbf{X})]^2\}, i = \overline{1, n}.$

Now, if we have in view the Definition 1.2, the Remark 1.3, respectively the relation (1.9a), then we obtain the relation

(2.32) $$d_2(\mathbf{Y}, \widehat{X}_0) = \left\|\mathbf{Y} - \widehat{X}_0\right\| = \sqrt{(\mathbf{Y} - \widehat{X}_0, \mathbf{Y} - \widehat{X}_0)} = [E(\left|\mathbf{Y} - \widehat{X}_0\right|^2)]^{1/2},$$

for the random vectors $\mathbf{Y}$ and $\widehat{X}_0 = g_0(\mathbf{X})$,as well as, the successive relations

2.32a $$d_2^2(\mathbf{Y}, \widehat{X}_0) = E(\left|\mathbf{Y} - \widehat{X}_0\right|^2) = E[(\mathbf{Y} - \widehat{X}_0)^2] =$$

$$= \left\|\mathbf{Y} - \widehat{X}_0\right\|^2 = (\mathbf{Y} - \widehat{X}_0, \mathbf{Y} - \widehat{X}_0) =$$

$$= \sum_{i=1}^{n} \left\|Y_i - \widehat{X}_0^{(i)}\right\|^2 =$$

$$= \sum_{i=1}^{n} E\left[\left(Y_i - \widehat{X}_0^{(i)}\right)^2\right] =$$

2.32b $$= \sum_{i=1}^{n} E\{[Y_i - g_0^{(i)}(\mathbf{X})]^2) =$$

$$= \sum_{i=1}^{n} E\{[Y_i - E[Y_i \mid \mathbf{X})]^2\} =$$

2.32c $$= \sum_{i=1}^{n} e_{\min}^{(i)}(Y_i, \widehat{X}_0^{(i)}) =$$

2.32d $$= e_{\min}(\mathbf{Y}, \widehat{X}_0),$$

which put in evidence just the equalities(2.29) and (2.29a).

In conclusion, the optimal estimator (2.28) is a nonlinear function that represents the conditional mean of the random vector $\mathbf{Y}$, then when the random vector $\mathbf{X}$ is given. Evidently, this optimal estimator $\widehat{X}_0$ is a random variable and its values are of the form

(2.33) $$\mathbf{M}(Y_i \mid X_j = x_j, j = \overline{1,n}) = \int_{-\infty}^{\infty} y_i f(y_i \mid x_1, x_2, ..., x_n) dy_j =$$

2.33a $$= \frac{1}{f(x_1, x_2, ..., x_n)} \int_{-\infty}^{\infty} y_i f(y_i, x_1, x_2, ..., x_n) dy_i, i = \overline{1,n},$$

for each point

(2.33b) $$\mathbf{x} = (x_1, x_2, .., x_n)^T \in \mathbb{D}_{\mathbf{x}} = \{\mathbf{x} \in \mathbb{R}^n | f(x_1, .., x_n) = f(\mathbf{x} > 0\},$$

if we had in view the following relations

(2.34) $$f(y_i \mid x_1, x_2, ..., x_n) = \frac{f(y_i, x_1, x_2, ..., x_n)}{f(x_1, x_2, ..., x_n)}, i = \overline{1,n}.$$

Therefore, for to solve a such problem of *the nonlinear estimation in the mean-square* we must to known the conditional densities of the forms (2.34). □

References

[1] Mihoc, I., Fătu, C. I., *Calculul probabilităţilor şi statistică matematică*, Casa de Editură Transilvania Pres, Cluj-Napoca, 2003

[2] Mihoc, I., Fătu, C. I., *Mean-square estimation and conditional densities*, RoGer 2004, The 6th Romanian-German Seminar on Approximation Theory and its Applications, Cluj-Napoca-Băişoara, Romania (to appear), June 3-June 6, 2004

[3] Mihoc, I., Fătu, C. I., *The orthogonality principle in the theory of statistical estimation*, 5 th Joint Conference on Mathematics and Computer Science, Debrecen, Hungary (to appear), June 9-12, 2004

[4] Rao, C. R., *Linear Statistical Inference and Its Applications*, John Wiley and Sons, Inc., New York, 1965

[5] Shiryaev, A. N., *Probability*, Springer-Verlag, New York Berlin, 1996

[6] Wilks, S. S., *Mathematical Statistics*, Wiley, New York, 1962

Christian University "Dimitrie Cantemir"
Faculty of Economics
Cluj-Napoca, Romania
E-mail address: cfatu@cantemir.cluj.astral.ro