CREATIVE MATH. & INF.
16 (2007), 27 - 35

Dedicated to Professor Ioan A. RUS on the occasion of his 70^{th} *anniversary*

About the orthogonal relations in the statistical estimation

ION MIHOC and 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 using a linear or nonlinear function of them. In this paper, using some definitions and properties of the linear mean-square estimation as well as the orthogonality principle we present some implications of them in the statistical estimation.

1. ORTHOGONALITY PRINCIPLE IN THE STATISTICAL ESTIMATION

Let X and Y two random variables and we suppose that only X can be observed. 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

$$E(\widehat{X}) = Y, \tag{1.1}$$

i.e., the estimator $\widehat{X}$ to be *unbiased.*

If $\widehat{X}$ is *an unbiased estimator*, then the matrix

$$\mathbf{K}_e = E[(\widehat{X} - E(\widehat{X}))(\widehat{X} - E(\widehat{X}))^T], \tag{1.2}$$

is its *covariance matrix* of which diagonal terms are the variances of the estimator's components. The estimator in this last case is called *the minimum variance unbiased estimator*. This type of estimator will be our choice for *the optimum* or *best estimator*.

Definition 1.1. We say that a function $X^* = g^*(X)$ *on* X is best estimator in the mean-square sense for the random variable Y if

$$E\{[Y - X^*]^2\} = E\{[Y - g^*(X)]^2\} = \inf_g E\{[Y - g(X)]^2\}. \tag{1.3}$$

Remark 1.1. In the next we consider $(n+1)$ random variables

$$Y, X_1, X_2, ..., X_n \in L^2(\Omega, K, P) \tag{1.4}$$

and we wish to estimate Y by *a nonlinear function* on random vector

$$\mathbf{X} = (X_1, X_2, ..., X_n)^T \tag{1.5}$$

Received: 15.09.2006. In revised form: 16.11.2006.
2000 *Mathematics Subject Classification.* 62H12, 62J12.
Key words and phrases. *Estimation, mean-square estimation, conditional means, orthogonality principle.*

of the form

$$\widehat{X}_0 = g_0(\mathbf{X}) = g_0(X_1, X_2, ..., X_n) \tag{1.6}$$

so as to minimize the mean-square error

$$e = e(Y, \widehat{X}_0) = E[(Y - \widehat{X}_0)^2], \tag{1.7}$$

that is, to have

$$\begin{aligned} e_{min}(Y, \widehat{X}_0) &= E[(Y - \widehat{X}_0)^2] = \\ &= E\left\{[Y - g_0(X_1, X_2, ..., X_n)]^2]\right\}. \end{aligned} \tag{1.8}$$

Theorem 1.1. [2] *The random variable*

$$\begin{aligned} \widehat{X}_0 &= g_0(X_1, X_2, ..., X_n) = g_0(\mathbf{X}) = \\ &= E[Y \mid (X_1, X_2, ..., X_n)^T] = \\ &= E(Y \mid \mathbf{X}), \end{aligned} \tag{1.9}$$

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

$$E[Y \mid \mathbf{X} = \mathbf{x}] = \int_{-\infty}^{\infty} y f(y \mid \mathbf{x}) dy, \tag{1.10}$$

for any $n-$dimensional real point $\mathbf{x}$ of the form

$$\mathbf{x} = (x_1, x_2, .., x_n)^T \in \mathbb{D}_{\mathbf{x}} = \{\mathbf{x} \in \mathbb{R}^n | f(x_1, .., x_n) = f(\mathbf{x}) > 0\}, \tag{1.11}$$

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

$$\begin{aligned} e_{min}(Y, \widehat{X}_0) &= \min_{\widehat{X}} E[(Y - \widehat{X})^2] = \\ &= E\left\{[Y - g_0(\mathbf{X})]^2]\right\} = \\ &= E\left\{[Y - E(Y \mid \mathbf{X})]^2\right\}. \end{aligned} \tag{1.12}$$

Remark 1.2. The estimation problem is considerably simplified if one seeks an estimate of X_0 by *a linear combination* of $X_1, X_2, ..., X_n$ of the form

$$g_n(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} a_i X_i, \ where\ a_i \in \mathbb{R},\ i = \overline{1, n}. \tag{1.13}$$

In a such case the problem is to find the values $\widehat{a}_1, \widehat{a}_2, ..., \widehat{a}_n$ of the constants $a_1, a_2, ..., a_n$ such that *the mean-square error*

$$e = e(a_1, a_2, ..., a_n) = M[(X_0 - \sum_{i=1}^{n} a_i X_i)^2] \tag{1.14}$$

is minimum, namely

$$\begin{aligned}
e_{\min} &= e_{\min}[X_0, g_n(X_1, X_2, ..., X_n)] = \\
&= \min_{a_i \in \mathbb{R}} E[(X_0 - \sum_{i=1}^{n} a_i X_i)^2] = \\
&= E[(X_0 - \underbrace{\sum_{i=1}^{n} \widehat{a}_i X_i}_{=\widehat{X}_0})^2] = \\
&= E[(X_0 - \widehat{X}_0)^2] = \\
&= e(\widehat{a}_1, \widehat{a}_2, ..., \widehat{a}_n).
\end{aligned} \tag{1.15}$$

If we consider that

$$E(X_i) = 0, \; i = \overline{1, n}, \tag{1.16}$$

then the constants $\widehat{a}_1, \widehat{a}_2, ..., \widehat{a}_n$ can be determined in terms of the second moments

$$K_{ij} = E(X_i X_j) = cov(X_i, X_j), \; i, j = \overline{0, n}. \tag{1.17}$$

Theorem 1.2. [3] *(The orthogonality principle) The constants* $\widehat{a}_i, i = \overline{1, n}$, *that minimize the mean-square error are such that the error vector*

$$X_0 - \widehat{X}_0 = X_0 - \sum_{i=1}^{n} \widehat{a}_i X_i, \tag{1.18}$$

is orthogonal to the random variables X_i, $i = \overline{1, n}$, *that is, if we have the following orthogonality relations*

$$E\left[(X_0 - \sum_{i=1}^{n} \widehat{a}_i X_i) X_j\right] = E\left[(X_0 - \widehat{X}_0) X_j\right] = 0, \; j = \overline{1, n}. \tag{1.19}$$

2. Relations implied by the orthogonality principle

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

$$L^p = L^p(\Omega, K, P) = \{X \in \mathcal{F}(\Omega, K, P) \mid E(|X|^p) < \infty\}, p \in \mathbb{N}^* \tag{2.1}$$

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

$$\beta_p = E(|X|^p) = \int_{\mathbb{R}} |x|^p \, dF(x) < \infty, p \in \mathbb{N}^*, \tag{2.2}$$

where

$$F(x) = P(X < x), x \in \mathbb{R} \tag{2.2a}$$

is the distribution function of the random variable X.

Lemma 2.1. [3] *The set* $L^p(\Omega, K, P)$ *represents a linear space.*

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

Definition 2.1. 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 as

$$d_2(X, Y) = \|X - Y\| = [E(|X - Y|^2)]^{1/2}. \tag{2.3}$$

Definition 2.2. 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

$$\begin{aligned} \underset{n\to\infty}{lim} d_2(X_n, X) &= \underset{n\to\infty}{lim} E(|X_n - X|^2)^{1/2} = \\ &= \underset{n\to\infty}{lim} E(|X_n - X|^2) = 0. \end{aligned} \tag{2.4}$$

We write

$$l.i.m.X_n = X \ or \ X_n \overset{m.p.}{\longrightarrow} X, n \to \infty, \tag{2.4a}$$

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

Remark 2.1. If $X \in L^2(\Omega, K, P)$, then

$$Var(X) = E[(X-m)^2] = E[|X - m|^2] = \|X - m\|^2 = d_2^2(X, m), \ m = E(X) \tag{2.5}$$

Remark 2.2. If we consider *the linear space* $H \equiv L^2(\Omega, K, P)$ and $X_i \in H, i = \overline{0, n}$, then the set

$$H_{n+1} = \{X \mid X = \sum_{i=0}^{n} a_i X_i, a_i \in \mathbb{R}, E(X_i^2) < \infty, i = \overline{0, n}\}, \tag{2.6}$$

is *a linear subspace of* H ($H_{n+1} \subset H$) generated by the random variables of the finite system

$$\{X_0, X_1, X_2, ..., X_n.\}$$

In a such case *the covariance matrix, denoted by* $\mathbf{K} = \mathbf{K_X}$, associated to the real random vector

$$\mathbf{X} = (X_0, X_1, X_2, ..., X_n)^T, \tag{2.7}$$

has the form

$$\begin{aligned} \mathbf{K} = \mathbf{K_X} &= E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \\ &= \begin{pmatrix} K_{00} & K_{01} & K_{02} & ... & K_{0n} \\ K_{10} & K_{11} & K_{12} & ... & K_{1n} \\ K_{20} & K_{21} & K_{22} & ... & K_{2n} \\ ... & ... & ... & ... & ... \\ K_{n0} & K_{n1} & K_{n2} & ... & K_{nn} \end{pmatrix}, \end{aligned} \tag{2.7a}$$

where

$$K_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)],\ i,j = \overline{0,n}, \tag{2.7b}$$
$$K_{ii} = E[(X_i - \mu_i)^2] = \sigma_i^2,\ \mu_i = E(X_i),\ i = \overline{0,n} \tag{2.7c}$$

and

$$\mu = E(\mathbf{X}) = (\mu_0, \mu_1, \mu_3, ..., \mu_n)^T. \tag{2.7d}$$

Remark 2.3. If $X', X'' \in H_{n+1}$ and

$$X' = \sum_{i=0}^{n} a_i' X_i, X'' = \sum_{i=0}^{n} a_i'' X_i, \tag{2.8}$$

then the scalar product associated

$$(X', X'') = E(X'X'') = \sum_{i=0}^{n}\sum_{j=0}^{n} a_i' a_j'' K_{ij} \tag{2.9}$$

has the properties

$$\begin{cases} a) & (X', X') \geq 0 \\ b) & (X', X') = 0 \Longleftrightarrow X' = 0 \\ c) & (aX' + b'', X''') = a(X', X''') + b(X'', X'''),\ a, b \in \mathbb{R} \end{cases} \tag{2.9a}$$

Also, using a such scalar product *the norm* will be defined as

$$\| X' \| = \sqrt{(X', X')}. \tag{2.9b}$$

More, *the distance in mean square* between X' and X'', that is,

$$\begin{aligned} d_2(X', X'') &= [E(|X' - X"|^2)]^{1/2} = \\ &= \sqrt{(X' - X'', X' - X'')} = \\ &= \|X' - X''\| \end{aligned} \tag{2.10}$$

is a *distance* in *the Euclidean subspace* H_{n+1}.

In accordance with the terminology of functional analysis, a space with the scalar product (2.9) is *a Hilbert space.* Hilbert space methods are extensively used in probability theory to study properties that depend only on the first two moments of random variables $(''L^2 - theory'')$.

Remark 2.4. If we consider *the linear subspace*

$$H_n = \{X \mid X = \sum_{i=1}^{n} a_i X_i,\ a_i \in \mathbb{R},\ E(X_i^2) < \infty,\ i = \overline{1,n}\},\ H_n \subset H_{n+1} \subset H, \tag{2.11}$$

then there is a random vector $\widehat{X}_0$ of the form

$$\widehat{X}_0 = \sum_{i=1}^{n} \widehat{a}_i X_i \tag{2.12}$$

which satisfies the following condition

$$\left\| X_0 - \widehat{X}_0 \right\| = \min_{\widehat{X} \in H_n} \left\| X_0 - \widehat{X} \right\| = \min_{a_i \in \mathbb{R}, i=\overline{1,n}} \left\| X_0 - \sum_{i=1}^{n} a_i X_i \right\|, \tag{2.13}$$

if the random variables $X_1, X_2, .., X_n$ are *linearly independent*, that is, the equation

$$P\left[\omega : \sum_{i=1}^{n} a_i X_i(\omega) = 0\right] = 1 \tag{2.14}$$

is satisfied only when all a_i, $i = \overline{1,n}$ are zero.

This last relation represents just the linearly independent of the random variables $X_1, X_2, .., X_n$ if we have in view the next theorem.

Theorem 2.1. *Let* $X \in L^2(\Omega, K, P)$. *If*

$$E(X^2) = 0, \tag{2.15}$$

then the random variable X take the value zero with probability one (*almost surely, almost everyhere), that is,*

$$P[\omega : X(\omega) = 0] = 1 \text{ or } P[\omega : X(\omega) \neq 0] = 0. \tag{2.16}$$

Proof. To prove this theorem we suppose that from the relation (2.15) would result an another relation of the form

$$P[\omega : X(\omega) \neq 0] \neq 0. \tag{2.17}$$

Then, a such relation implies the existence of a real number $\eta > 0$ such that

$$P[\omega : |X(\omega)| > \eta] = \int\limits_{|x|>\eta} f(x)dx \neq 0. \tag{2.17a}$$

But, if we have in view the definition for the second moment $\alpha_2 = E(X^2)$, we obtain

$$\begin{aligned} E(X^2) &= \int\limits_{-\infty}^{\infty} x^2 f(x)dx = \\ &= \int\limits_{|x|\leq\eta} x^2 f(x)dx + \int\limits_{|x|>\eta} x^2 f(x)dx = \\ &\geq \int\limits_{|x|>\eta} x^2 f(x)dx \geq \eta^2 \int\limits_{|x|>\eta} f(x)dx > 0, \end{aligned}$$

and, from here, it follows the inequality

$$E(X^2) > 0, \tag{2.15a)}$$

which evidently represents a contradiction of the hypothesis (2.15).

Theorem 2.2. *The orthogonality relations from the Theorem 1.2 , that is,*

$$E\left[(X_0-\sum_{i=1}^{n}\widehat{a}_iX_i)X_j\right]=E\left[(X_0-\widehat{X}_0)X_j\right]=0,\ j=\overline{1,n}. \tag{2.18}$$

are equivalently with the relation

$$(X_0-\widehat{X}_0,\widehat{X})=0, \tag{2.19}$$

or with the relations

$$\sum_{i=1}^{n}\widehat{a}_i(X_i,X_j)=(X_0,X_j),j=\overline{1,n}, \tag{2.20}$$

where $X_0,X_1,X_2,...,X_n\in L^2(\Omega,K,P)=H$ *and*

$$\widehat{X}_0=\sum_{i=1}^{n}\widehat{a}_iX_i,\widehat{a}_i\in\mathbb{R},\ i=\overline{1,n},\widehat{X}_0\in H_n,H_n\subset H \tag{2.20a}$$

$$\widehat{X}=\sum_{j=1}^{n}a_jX_j,a_j\in\mathbb{R},j=\overline{1,n},\ \widehat{X}\in H_n,H_n\subset H. \tag{2.20b}$$

Proof. Indeed, using the relations (2.18) and (2.19) it follows a new relation

$$M\left[(X_0-\widehat{X}_0)\widehat{X}\right]=0, \tag{2.21}$$

which, then when we have in view the properties of the scalar product, can be represented in the form

$$(\widehat{X}_0,\widehat{X})=(X_0,\widehat{X}). \tag{2.22}$$

Then, using the relations (2.20a) and (2.20b), the scalar products $(\widehat{X}_0,\widehat{X})$ and $(X_0,\widehat{X})$ can be written as

$$\begin{aligned}(\widehat{X}_0,\widehat{X})&=(\widehat{X}_0,\sum_{j=1}^{n}a_jX_j)=\sum_{j=1}^{n}a_j(\widehat{X}_0,X_j)=\\&=\sum_{j=1}^{n}a_j\left(\sum_{i=1}^{n}\widehat{a}_iX_i,X_j\right)=\\&=\sum_{j=1}^{n}a_j\left[\sum_{i=1}^{n}\widehat{a}_i(X_i,X_j)\right],\end{aligned} \tag{2.22a}$$

$$(X_0,\widehat{X})=(X_0,\sum_{j=1}^{n}a_jX_j)=\sum_{j=1}^{n}a_j[(X_0,X_j)]. \tag{2.22b}$$

In final, using these two representations of the scalar products, from (2.22) we obtain the relations (2.20) which, in fact, represents an another form of the linear system

$$\begin{cases} K_{11}\widehat{a}_1 & + & K_{12}\widehat{a}_2 & + & ... & + & K_{1n}\widehat{a}_n & = & K_{01} \\ K_{21}\widehat{a}_1 & + & K_{22}\widehat{a}_2 & + & ... & + & K_{2n}\widehat{a}_n & = & K_{02} \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ K_{n1}\widehat{a}_1 & + & K_{n2}\widehat{a}_2 & + & ... & + & K_{nn}\widehat{a}_n & = & K_{0n} \end{cases} \tag{2.23}$$

Remark 2.5. The matrix form of this system will be

$$\begin{bmatrix} (X_1,X_1) & (X_1,X_2) & ... & (X_1,X_n) \\ (X_2,X_1) & (X_2,X_2) & ... & (X_2,X_n) \\ ... & ... & ... & ... \\ (X_n,X_1) & (X_n,X_2) & ... & (X_n,X_n) \end{bmatrix} \begin{bmatrix} \widehat{a}_1 \\ \widehat{a}_2 \\ ... \\ \widehat{a}_n \end{bmatrix} = \begin{bmatrix} (X_0,X_1) \\ (X_0,X_2) \\ ... \\ (X_0,X_n) \end{bmatrix}, \tag{2.24}$$

where

$$G = \begin{bmatrix} (X_1,X_1) & (X_1,X_2) & ... & (X_1,X_n) \\ (X_2,X_1) & (X_2,X_2) & ... & (X_2,X_n) \\ ... & ... & ... & ... \\ (X_n,X_1) & (X_n,X_2) & ... & (X_n,X_n) \end{bmatrix} \tag{2.25}$$

is Gram's matrix associated to the $n-$dimensional random vector

$$\mathbf{X} = (X_1, X_2, ..., X_n)^T,$$

where

$$\begin{cases} (X_i, X_j) & = & E(X_iX_j) = cov(X_i, X_j), & i,j = \overline{1,n}, i \neq j \\ (X_i, X_i) & = & E(X_i^2) = \sigma_i^2, & i = \overline{1,n}, \end{cases} \tag{2.25a}$$

then when we have in view the hypothesis

$$E(X_i) = 0, i = \overline{1,n}. \tag{2.25b}$$

If the components of the random vector $\mathbf{X} = (X_1, X_2, ..., X_n)^T$ are *mutually independent* , then from (2.25b) it follows that these components will be and *mutually orthogonal* , that is, we have

$$\begin{cases} (X_i, X_j) & = & M(X_iX_j) = cov(X_i, X_j) = 0, & i,j = \overline{1,n}, i \neq j \\ (X_i, X_i) & = & M(X_i^2) = \sigma_i^2, & i = \overline{1,n}. \end{cases} \tag{2.25c}$$

In these new conditions *the determinant of Gram* satisfies the condition

$$det\ G = \prod_{i=1}^{n} \sigma_i^2 \neq 0, \tag{2.26}$$

which implies the following conclusion: the random variables $X_1, X_2, ..., X_n$ are *linearly independent* and the unique solution of the system (2.23) will be

$$\widehat{a}_i = \frac{(X_0, X_i)}{\sigma_i^2},\ i = \overline{1,n}. \tag{2.27}$$

Definition 2.3. Let $\mathcal{F}(\Omega, K, P)$ be the family of all random variables defined on (Ω, K, P) and

$$H = L^2(\Omega, K, P) = \left\{ X \in \mathcal{F}(\Omega, K, P) \mid E(|X|^2) < \infty \right\}, \tag{2.28}$$

be the set of random variables with finite moments of order 2, that is,

$$\beta_2 = E(|X|^2) = \int_{\mathbb{R}} |x|^2 \, dF(x) < \infty. \tag{2.29}$$

Then, the random variable

$$\widehat{X}_0 = \sum_{i=1}^{n} \widehat{a}_i X_i, \ \widehat{X}_0 \in H_n, \tag{2.30}$$

represents an optimal estimator (the best estimator) for the unknown random variable $X_0, X_0 \in H_{n+1}$, if

$$\left\| X_0 - \widehat{X}_0 \right\| = \left\| X_0 - \sum_{i=1}^{n} \widehat{a}_i X_i \right\| =$$

$$\min_{\widehat{X} \in H_n} \left\| X_0 - \widehat{X} \right\| = \min_{a_i \in \mathbb{R}, i=\overline{1,n}} \left\| X_0 - \sum_{i=1}^{n} a_i X_i \right\|. \tag{2.31}$$

The geometric meaning of the optimal estimator $\widehat{X}_0$, $\widehat{X}_0 \in H_n$, is the following: $\widehat{X}_0$ is *the "projection"* of $X_0, X_0 \in H_{n+1}$ on *the linear subspace* H_n.

REFERENCES

[1] Fătu Cristina Ioana, *The orthogonality principle and conditional densities*, Carpathian Journal of Mathematics, Vol. 20, No. 1, pp. 31-38, 2004

[2] Fătu Cristina Ioana, *Metode optimale în teoria estimării statistice,* Editura Risoprint, Cluj-Napoca, 2005

[3] Mihoc I. and Fătu Cristina Ioana, *Some implication of the orthogonal principle in the statistical estimation,* International Conference on Numerical Analysis and Approximation Theory, Cluj-Napoca, Romania, July 5-8, 2006 (to appear)

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

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

FACULTY OF ECONOMICS AT CHRISTIAN UNIVERSITY "DIMITRIE CANTEMIR"
AL. V. VOEVOD 47
400592 CLUJ-NAPOCA, ROMANIA
E-mail address: imihoc@cantemir.cluj.astral.ro
E-mail address: cristinafatu@cantemir.cluj.astral.ro