Mathematical Problems of Computer Science 28, 2007, 114–119.

On Identification of Distributions of Two Independent Objects

Evgueni A. Haroutunian and Parandzem M. Hakobyan

Institue for Informatics and Automation Problems of NAS of RA
e-mail: evhar@ipia.sci.am

Abstract

Ahlswede and Haroutunian formulated new problems on multiple hypotheses testing and on identification of hypotheses. The problems of identification of distribution and of distributions ranking for one object were solved completely. In this paper we study the problem of identification of distributions for two independent objects.

1 Introduction

In [1], [2] Ahlswede and Haroutunian formulated an ensemble of new problems on multiple hypotheses testing for many objects and on identification of hypotheses. Noted problems are extentions of those investigated in the books [3] and [4]. In papers [5] and [6] the problem of multiple hypotheses testing for many objects which independently follow to one of given $M(\geq 2)$ probability distributions (PDs) is solved. The problem is a generalization of those investigated in [7] for testing of many hypotheses concerning one object. The problems of identification of distribution and of distributions ranking for one object were solved in [2] entirely. In this paper we study the problem of identification of distributions for two independent objects.

Let us recall main definitions from [5] and [7]. $\mathcal{P}(\mathcal{X})$ is the space of all PDs on finite set $\mathcal{X}$. There are given M PDs $G_m \in \mathcal{P}(\mathcal{X})$, $m = \overline{1,M}$. The random variable (RV) X taking values on $\mathcal{X}$ follows to one of M PDs G_m, $m = \overline{1,M}$. The statistician must accept one of M hypotheses $H_l : G = G_l$, $l = \overline{1,M}$, on the base of a sequence of results of N observations of the object $\mathbf{x} = (x_1, ..., x_n, ..., x_N)$, $x_n \in \mathcal{X}$, $n = \overline{1,N}$. The procedure of decision making is a non-randomized test $\varphi_N(\mathbf{x})$, which can be defined by partition of the sample space $\mathcal{X}^N$ on M disjoint subsets $\mathcal{A}_l^N = \{\mathbf{x} : \varphi_N(\mathbf{x}) = l\}$, $l = \overline{1,M}$. The set $\mathcal{A}_l^N$ contains all vectors $\mathbf{x}$ for which the hypothesis H_l is adopted. The probability $\alpha_{l|m}(\varphi_N)$ of the erroneous acceptance of hypothesis H_l provided that H_m is true, is equal to $G_m^N(A_l^N)$, $l \neq m$. The probability to reject H_m, when it is true, is

$$\alpha_{m|m}(\varphi_N) \triangleq \sum_{l \neq m} \alpha_{l|m}(\varphi_N). \tag{1}$$

The error probability exponents, called "reliabilities" of the infinite sequence of tests φ, are defined as

$$E_{l|m}(\varphi) \triangleq \varliminf_{N \to \infty} -\frac{1}{N} \log \alpha_{l|m}(\varphi_N), \quad m, l = \overline{1,M}. \tag{2}$$

It follows from (1) that

$$E_{m|m}(\varphi) = \min_{l \neq m} E_{l|m}(\varphi), \quad m = \overline{1,M}. \tag{3}$$

The matrix $\mathbf{E}(\varphi) = \{E_{l|m}(\varphi)\}$ is the reliability matrix of the sequence φ of tests. It was studied in [7].

The sequence of tests φ^* is called logarithmically asymptotically optimal (LAO) if for given positive values of $M-1$ diagonal elements of the matrix $\mathbf{E}(\varphi^*)$ maximal values to all other elements of it are provided.

Now let us consider the model with two objects. Let X_1 and X_2 be independent RV taking values in the same finite set $\mathcal{X}$ with one of M PDs, they are characteristics of corresponding independent objects. The random vector (X_1, X_2) assumes values $(x^1, x^2) \in \mathcal{X} \times \mathcal{X}$.

Let $(\mathbf{x}_1, \mathbf{x}_2) = ((x_1^1, x_1^2), ..., (x_n^1, x_n^2), ..., (x_N^1, x_N^2))$, $x_n^i \in \mathcal{X}$, $i = \overline{1,2}$, $n = \overline{1,N}$, be a posequence of results of N independent observations of the vector (X_1, X_2). The statistician must define unknown PDs of the objects on the base of observed data. The selection for each object must be made from the same known set of hypotheses: $H_m : G = G_m$, $m = \overline{1,M}$. We call the procedure of making decision on the base of N pairs of observations the test for two objects and denote it by Φ_N. Because of the objects independence test Φ_N may be considered as the pair of the tests φ_N^1 and φ_N^2 for the respective separate objects. We shall denote the whole compound test sequence by Φ.

Let $\alpha_{l_1,l_2|m_1,m_2}(\Phi_N)$ be the probability of the erroneous acceptance by the test Φ_N of the hypotheses pair (H_{l_1}, H_{l_2}) provided that the pair (H_{m_1}, H_{m_2}) is true, where $(m_1, m_2) \neq (l_1, l_2)$, $m_i, l_i = \overline{1,M}$, $i = 1,2$. The probability to reject a true pair of hypotheses (H_{m_1}, H_{m_2}) by analogy with (1) is the following:

$$\alpha_{m_1,m_2|m_1,m_2}(\Phi_N) \triangleq \sum_{(l_1,l_2) \neq (m_1,m_2)} \alpha_{l_1,l_2|m_1,m_2}(\Phi_N). \tag{4}$$

Corresponding limits $E_{l_1,l_2|m_1,m_2}(\Phi)$ of the error probability exponents of the sequence of tests Φ, are also called reliabilities:

$$E_{l_1,l_2|m_1,m_2}(\Phi) \triangleq \overline{\lim_{N \to \infty}} -\frac{1}{N} \log \alpha_{l_1,l_2|m_1,m_2}(\Phi_N), \quad m_i, l_i = \overline{1,M}, \quad i = 1,2. \tag{5}$$

As in (3) it follows from (5) that

$$E_{m_1,m_2|m_1,m_2}(\Phi) = \min_{(l_1,l_2) \neq (m_1,m_2)} E_{l_1,l_2|m_1,m_2}(\Phi). \tag{6}$$

We shall call the test sequence Φ^* LAO for the model with two objects if for given positive values of certain $2(M-1)$ elements of the reliability matrix $\mathbf{E}(\Phi^*)$ the procedure provides maximal values for all other elements of it.

In the Section 2 we call to mind result of [7], then in Section 3 result from [2] on identification of PD for one object will be formulated and in Section 4 the problem of identification of PDs for two objects will be solved.

LAO Testing of Hypotheses for One Object

We remind the results of paper [7] for further use. We need the divergence (Kullback-Leibler distance) $D(Q||G)$ of PDs $Q, G \in \mathcal{P}(\mathcal{X})$, defined as usual (see [8]):

$$D(Q||G) = \sum_{x \in \mathcal{X}} Q(x) \log \frac{Q(x)}{G(x)}.$$

For given positive elements $E_{1|1}, E_{2|2}, \ldots, E_{M-1|M-1}$ we can divide $\mathcal{P}(\mathcal{X})$ on M subsets.

$$\mathcal{R}_l \triangleq \{Q : D(Q||G_l) \le E_{l|l}\}, \quad l = \overline{1, M-1}, \tag{7.a}$$

$$\mathcal{R}_M \triangleq \{Q : D(Q||G_l) > E_{l|l}, \quad l = \overline{1, M-1}\} = \mathcal{P}(\mathcal{X}) - \bigcup_{l=1}^{M-1} \mathcal{R}_l. \tag{7.b}$$

and consider the following values:

$$E^*_{l|l} = E^*_{l|l}(E_{l|l}) \triangleq E_{l|l}, \quad l = \overline{M-1}, \tag{8.a}$$

$$E^*_{l|m} = E^*_{l|m}(E_{l|l}) \triangleq \inf_{Q \in \mathcal{R}_l} D(Q||G_m), \quad m = \overline{1, M}, \ m \ne l, \ l = \overline{1, M-1}, \tag{8.b}$$

$$E^*_{M|m} = E^*_{M|m}(E_{1|1}, \ldots, E_{M-1|M-1}) \triangleq \inf_{Q \in \mathcal{R}_M} D(Q||G_m), \quad m = \overline{1, M-1}, \tag{8.c}$$

$$E^*_{M|M} = E^*_{M|M}(E_{1|1}, \ldots, E_{M-1|M-1}) \triangleq \min_{l=\overline{1,M-1}} E^*_{l|M}. \tag{8.d}$$

The main result of paper [7] is:

Theorem 1: If the distributions G_m, $m = \overline{1, M}$, are different, that is all elements of the matrix $\{D(G_l||G_m)\}$, are strictly positive, then two statements hold:

a) when the given numbers $E_{1|1}, E_{2|2}, \ldots, E_{M-1|M-1}$ satisfy conditions

$$0 < E_{1|1} < \min_{l=\overline{2,M}} D(G_l||G_1), \tag{9.a}$$

$$0 < E_{m|m} < \min\left[\min_{l=\overline{1,m-1}} E^*_{l|m}(E_{l|l}), \min_{l=\overline{m+1,M}} D(G_l||G_m)\right], \quad m = \overline{2, M-1}, \tag{9.b}$$

then there exists a LAO sequence of tests φ^*, the reliability matrix of which $\mathbf{E}(\varphi^*) = \{E^*_{l|m}\}$ is defined in (8) and all elements of it are strictly positive;

b) even if one of conditions (9) is violated, then the reliability matrix of any such test includes at least one element equal to zero (that is the corresponding error probability does not tend to zero exponentially).

Corollary 1[5]: If in contradiction to conditions (9) one or several element $E_{m|m}$, $m \in [1, M-1]$, of the reliability matrix are equal to zero, then the elements of the matrix determined in functions of this $E_{m|m}$ will be given as in the case of Stain's lemma [8]

$$E'_{m|l}(E_{m|m}) = D(G_m||G_l), \quad l = \overline{1, M}, \ l \ne m,$$

and the remaining elements of the matrix $\mathbf{E}(\varphi^*)$ are defined by $E_{l|l} > 0$, $l \ne m$, $l = \overline{1, M-1}$, as follows from Theorem1:

$$E'_{l|k} = \inf_{Q: D(Q||G_l) \le E_{l|l}} D(Q||G_k),$$

$$E'_{M|k} = \inf_{Q: D(Q||G_i) > E_{i|i}, i = \overline{1,M-1}} D(Q||G_k).$$

Identification of the Probability Distribution of an Object

First it is necessary to formulate our meaning of the identification problem for one object, which was considered and solved in [1] and [2]. We have one object, and there are known $M \geq 2$ possible PDs.

What is the identification? It is the answer to the question whether r-th distribution occured, or not, as in the testing problem, must this answer be given on the base of a sample $\mathbf{x}$ and a test $\varphi_N^*(\mathbf{x})$.

There are two error probabilities for each $r \in [1, M]$: the probability $\alpha_{l\neq r|m=r}(\varphi_N)$ to accept l different from r, when r is in reality, and the probability $\alpha_{l=r|m\neq r}(\varphi_N)$ that r is accepted, when it is not correct.

The probability $\alpha_{l\neq r|m=r}(\varphi_N)$ is already known, it coincides with the probability $\alpha_{r|r}(\varphi_N)$ which is equal to $\sum_{l:l\neq r} \alpha_{l|r}(\varphi_N)$. The corresponding reliability $E_{l\neq r|m=r}(\varphi)$ is equal to $E_{r|r}(\varphi)$ which satisfies the equality (3).

And what is the reliability approach to identification? It is necessary to determine the optimal dependence of $E^*_{l=r|m\neq r}$ upon given $E^*_{l\neq r|m=r} = E^*_{r|r}$, which can be assigned value satisfying conditions (9).

The result from paper [2] is:

Theorem 2: In the case of distinct PDs $G_1, G_2, ..., G_M$, for a given sample $\mathbf{x}$ we define as type Q, and when $Q \in \mathcal{R}_t^{(N)}$ we accept the hypothesis r. Under condition that the probabilities of all M hypotheses are positive the reliability of such test $E_{l=r|m\neq r}$ for given $E_{l\neq r|m=r} = E_{r|r}$ is the following:

$$E_{l=r|m\neq r}(E_{r|r}) = \min_{m:m\neq r} \inf_{Q:D(Q\|G_r)\leq E_{r|r}} D(Q\|G_m), \quad r \in [1, M].$$

Identification of the Probability Distributions of Two Independent Objects.

We begin by main results from [6] for two independent objects and M hypotheses testing concerning each of them.

Lemma: If elements $E_{l|m}(\varphi^i)$, $m, l = \overline{1, M}$, $i = 1, 2$, are strictly positive, then the following equalities hold for $\Phi = (\varphi^1, \varphi^2)$:

$$E_{l_1,l_2|m_1,m_2}(\Phi) = \sum_{i=1}^{2} E_{l_i|m_i}(\varphi^i), \text{ if } m_1 \neq l_1, \ m_2 \neq l_2, \tag{10.a}$$

$$E_{l_1,l_2|m_1,m_2}(\Phi) = E_{l_i|m_i}(\varphi^i), \text{ if } m_{3-i} = l_{3-i}, \ m_i \neq l_i, \ i = 1, 2. \tag{10.b}$$

The LAO test Φ^* is the compound test consisting of the pair of LAO tests $\varphi^{*,1}$ and $\varphi^{*,2}$ for respective separate objects, and for it the equalities (10.a) and (10.b) take place. The statistician have to answer the question whether the pair of distributions (r_1, r_2) occurred or not. Let us consider two types of error probabilities for each pair (r_1, r_2), $r_1, r_2 \in [1, M]$. We denote by $\alpha^N_{(l_1,l_2)\neq(r_1,r_2)|(m_1,m_2)=(r_1,r_2)}$ the probability, that pair (r_1, r_2) is true, but it is rejected. Note that this probability is equal to $\alpha_{r_1,r_2|r_1,r_2}(\Phi_N)$. Let $\alpha^N_{(l_1,l_2)=(r_1,r_2)|(m_1,m_2)\neq(r_1,r_2)}$ be the probability that (r_1, r_2) is accepted, when it is not correct. The corresponding reliabilities are $E_{(l_1,l_2)\neq(r_1,r_2)|(m_1,m_2)=(r_1,r_2)} = E_{r_1,r_2|r_1,r_2}$ and $E_{(l_1,l_2)=(r_1,r_2)|(m_1,m_2)\neq(r_1,r_2)}$. Our aim is to determine the dependence of $E_{(l_1,l_2)=(r_1,r_2)|(m_1,m_2)\neq(r_1,r_2)}$ on given $E_{r_1,r_2|r_1,r_2}(\Phi_N)$.

Now let us suppose that hypotheses $G_1, G_2, ..., G_M$ have a priori positive probabilities $\Pr(r)$, $r = \overline{1,M}$, and consider the probability, which we are interested:

$$\alpha^N_{(l_1,l_2)=(r_1,r_2)|(m_1,m_2)\neq(r_1,r_2)} = \frac{\Pr^N((m_1,m_2)\neq(r_1,r_2),(l_1,l_2)=(r_1,r_2))}{\Pr((m_1,m_2)\neq(r_1,r_2))} =$$

$$= \frac{\sum\limits_{(m_1,m_2):(m_1,m_2)\neq(r_1,r_2)} \alpha_{(m_1,m_2)|(r_1,r_2)}\Pr((m_1,m_2))}{\sum\limits_{(m_1,m_2)\neq(r_1,r_2)} \Pr(m_1,m_2)}.$$

Consequently, we obtain that

$$E_{(l_1,l_2)=(r_1,r_2)|(m_1,m_2)\neq(r_1,r_2)} = \min_{(m_1,m_2):(m_1,m_2)\neq(r_1,r_2)} E_{r_1,r_2|m_1,m_2}. \quad (11)$$

For every LAO tests Φ^* from (6), (10) and (11) we obtain that

$$E_{(l_1,l_2)=(r_1,r_2)|(m_1,m_2)\neq(r_1,r_2)} = \min_{m_1\neq r_1, m_2\neq r_2} \left(E^I_{r_1|m_1}(E_{r_1|r_1}), E^{II}_{r_2|m_2}(E_{r_2|r_2})\right), \quad (12)$$

where $E^I_{r_1|m_1}(E_{r_1|r_1}), E^{II}_{r_2|m_2}(E_{r_2|r_2})$ are determined by (8) for, correspondingly, the first and the second objects. For every LAO test Φ^* from (6) and (10) we deduce that

$$E_{r_1,r_2|r_1,r_2} = \min_{m_1\neq r_1, m_2\neq r_2} \left(E^I_{r_1|m_1}, E^{II}_{r_2|m_2}\right) = \min\left(E^I_{r_1|r_1}, E^{II}_{r_2|r_2}\right). \quad (13)$$

and each of $E^I_{r_1|r_1}, E^{II}_{r_2|r_2}$ satisfy the following conditions (see Theorem 1, condition (9)):

$$0 < E^I_{r_1|r_1} < \min\left[\min_{l=\overline{1,r_1-1}} E^*_{l|m}(E^I_{l|l}), \min_{l=\overline{r_1+1,M}} D(G_l||G_{r_1})\right], \quad (14.a)$$

$$0 < E^{II}_{r_2|r_2} < \min\left[\min_{l=\overline{1,r_2-1}} E^*_{l|m}(E^{II}_{l|l}), \min_{l=\overline{r_2+1,M}} D(G_l||G_{r_2})\right]. \quad (14.b)$$

From (8.b) we see that the elements $E^*_{l|m}(E^I_{l|l})$, $l = \overline{1,r_1-1}$ and $E^*_{l|m}(E^{II}_{l|l})$, $l = \overline{1,r_2-1}$ are determined only by $E^I_{l|l}$ and $E^{II}_{l|l}$. But we are considering only elements $E^I_{r_1|r_1}$ and $E^{II}_{r_2|r_2}$. We can use Corollary 1 and apper estimate (14.a) and (14.b) as follows:

$$0 < E^I_{r_1|r_1} < \min\left[\min_{l=\overline{1,r_1-1}} D(G_{r_1}||G_l), \min_{l=\overline{r_1+1,M}} D(G_l||G_{r_1})\right], \quad (15.a)$$

$$0 < E^{II}_{r_2|r_2} < \min\left[\min_{l=\overline{1,r_2-1}} D(G_{r_2}||G_l), \min_{l=\overline{r_2+1,M}} D(G_l||G_{r_2})\right]. \quad (15.b)$$

Let us denote $r = \max(r_1, r_2)$ and $k = \min(r_1, r_2)$. From (13) we have that, when $E_{r_1,r_2|r_1,r_2} = E^I_{r_1|r_1}$, then $E^I_{r_1|r_1} \leq E^{II}_{r_2|r_2}$ and when $E_{r_1,r_2|r_1,r_2} = E^{II}_{r_2|r_2}$, then $E^{II}_{r_2|r_2} \leq E^I_{r_1|r_1}$. Hence, it can be implied that given strictly positive elemen $E_{r_1,r_2|r_1,r_2}$ must meet both inequalities (15.a) and (15.b), and the combination of these restrictions gives

$$E_{r_1,r_2|r_1,r_2} < \min\left[\min_{l=\overline{1,r-1}} D(G_r||G_l), \min_{l=\overline{k+1,M}} D(G_l||G_k)\right]. \quad (16)$$

using (14) we can determine reliability $E_{(l_1,l_2)=(r_1,r_2)|(m_1,m_2)\neq(r_1,r_2)}$ in function of $E_{r_1,r_2|r_1,r_2}$ as follows:

$$E_{(l_1,l_2)=(r_1,r_2)|(m_1,m_2)\neq(r_1,r_2)}\left(E_{r_1,r_2|r_1,r_2}\right) =$$

$$= \min_{m_1\neq r_1, m_2\neq r_2}\left[E_{r_1|m_1}(E_{r_1,r_2|r_1,r_2}), E_{r_2|m_2}(E_{r_1,r_2|r_1,r_2})\right], \qquad (17)$$

where $E_{r_1|m_1}(E_{r_1,r_2|r_1,r_2})$ and $E_{r_2|m_2}(E_{r_1,r_2|r_1,r_2})$ are determined by (8.b).

Finally we obtained

Theorem 3: If the distributions G_m, $m = \overline{1,M}$, are different and the given strictly positive number $E_{r_1,r_2|r_1,r_2}$ satisfy condition (16), then the reliability $E_{(l_1,l_2)=(r_1,r_2)|(m_1,m_2)\neq(r_1,r_2)}$ is defined in (17).

References

[1] E. A. Haroutunian, "Reliability in multiple hypotheses testing and identification problems". Proceedings of the NATO ASI, Yerevan, 2003. NATO Science Series III: Computer and Systems Sciences – vol. 198, pp. 189–201. IOS Press, 2005.

[2] R. F. Ahlswede and E. A. Haroutunian, "On logarithmically asymptotically optimal testing of hypotheses and identification". Lecture Notes in Computer Science, vol. 4123," General Theory of Information Transfer and Combinatorics" Springer, pp. 462–478, 2006.

[3] R. E. Bechhofer, J. Kiefer, and M. Sobel, *Sequential identification and ranking procedures.* The University of Chicago Press, Chicago, 1968.

[4] R. F. Ahlswede and I. Wegener, *Search problems.* Wiley, New York, 1987.

[5] E. A. Haroutunian and P. M. Hakobyan, "On LAO testing of multiple hypothesis for pair of objects", *Mathematical Problems of Computer Science vol. 26*, pp. 92–100, 2005.

[6] E. A. Haroutunian and P. M. Hakobyan, "On multiple hypotheses LAO testing for many independent objects". (Acceptted to IEEE International Symposium on Information Theory, France, Nice, 2007).

[7] E. A. Haroutunian, "Logarithmically asymptotically optimal testing of multiple statistical hypotheses", *Problems of Control and Information Theory*, vol. 19(5-6), pp. 413–421, 1990.

[8] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*, Academic Press, New York, 1981.

Երկու անկախ օբյեկտների բաշխումների նույնականացման մասին

Ե. Ա. Հարությունյան և Փ. Մ. Հակոբյան

Ամփոփում

Ալսվեղեն և Հարությունյանը լուծել են վարկածների նույնականացման խնդիրը մեկ օբյեկտի դեպքում:

Այս հոդվածը նվիրված է բաշխումների նույնականացման խնդրի լուծմանը երկու անկախ օբյեկտներից կազմված մոդելի համար: