ASSIGNMENT OF ARBITRARILY DISTRIBUTED RANDOM SAMPLES TO THE FIXED PROBABILITY DISTRIBUTION AND ITS RISK

E.E. Zhuk[1], D.D. Dus[2]
Belarusian State University
Minsk, BELARUS
e-mail: [1]zhukee@mail.ru, [2]dzianisdus@gmail.com

Abstract

The problem of statistical assignment of arbitrarily distributed random samples to the fixed probability distribution is considered. The decision rule based on the maximum likelihood method is proposed and its efficiency is analytically examined. The case of two samples of the same size and the Fisher model is studied.

1 Introduction

Let $m \geq 2$ random samples $X^{(1)}, \ldots, X^{(m)}$ be determined in the observation space R^N ($N \geq 1$) and the following conditions be satisfied.

1. Each sample $X^{(i)} = \{x_t^{(i)}\}_{t=1}^{n_i}$ consists of independent and identically distributed random vectors $x_t^{(i)} \in R^N$, $t = \overline{1, n_i}$ (n_i is the sample size) with the same probability density $p_i(x)$:

$$p_i(x) \geq 0, \quad x \in R^N : \quad \int_{R^N} p_i(x)dx = 1, \quad i = \overline{1, m}. \tag{1}$$

2. Samples $X^{(1)}, \ldots, X^{(m)}$ are independent in total.

Suppose that all densities $\{p_i(x)\}_{i=1}^m$ from (1) are unknown and distinguished from the fixed probability density function, which is often referred as hypothetical density function [1, 2]:

$$p(x) \geq 0, \quad x \in R^N : \quad \int_{R^N} p(x)dx = 1. \tag{2}$$

The problem is to choose the one of samples $\{X^{(i)}\}_{i=1}^m$ that is closer to the hypothetical density (2) in terms of the distribution simillarity.

Note, that the declared problem differs from so-called "goodness of fit testing" problem [1, 2]: samples $\{X^{(i)}\}_{i=1}^m$ are obtained from corresponding probability densities (1), but not from the hypothetical density (2). Also the problem differs from the classification problem [3, 4]: there is the only one class, determined by the density (2), to which one of samples $\{X^{(i)}\}_{i=1}^m$ should be assigned.

The problem is to construct the decision rule (DR):

$$d = d(X^{(1)}, \ldots, X^{(m)}) \in M, \quad M = \{1, \ldots, m\}, \tag{3}$$

to solve the specified assignment problem.

2 Maximum likelihood method and its risk

As it earlier was proposed in [4], the maximum likelihood method [1, 2, 3, 4] can be used to solve the assignment problem:

$$d = d(X^{(1)}, \ldots, X^{(m)}) = \arg\max_{i \in M} P(X^{(i)}); \tag{4}$$

$$P(X^{(i)}) = \prod_{t=1}^{n_i} p(x_t^{(i)}), \quad i \in M,$$

where $P(X^{(i)})$ is the hypothetical likelihood function [1, 2] evaluated for the sample $X^{(i)}$.

Theorem. *Let the following integrals be finite:*

$$\int_{R^N} |\ln(p(x))| p_i(x) dx < +\infty, \quad i \in M, \tag{5}$$

where $\{p_i(x)\}_{i \in M}$, $p(x)$ *are densities from (1), (2).*

If for values

$$H_i = H(p_i(\cdot), p(\cdot)) = \int_{R^N} \ln(p(x)) p_i(x) dx, \quad i \in M, \tag{6}$$

the condition

$$\exists d^0 \in M : H_{d^0} > H_i, \forall i \neq d^0, \quad i \in M,$$

is satisfied, and all samples $\{X^{(i)}\}_{i=1}^m$ *have the same size:*

$$n_i = n, \quad i \in M, \tag{7}$$

then for the decision rule (4) the following statement is true:

$$d = d(X^{(1)}, \ldots, X^{(m)}) \xrightarrow{a.s.} d^0, \quad n \to +\infty; \tag{8}$$

$$d^0 = \arg\max_{i \in M} H_i.$$

Analytical results described above allow us to introduce the generalization of the traditional risk (like as in [4]) as the measure of efficiency of the decision rule (4):

$$r = r(d(X^{(1)}, \ldots, X^{(m)})) = P\{d(X^{(1)}, \ldots, X^{(m)}) \notin D^0\}; \tag{9}$$

$$D^0 = \{k: \ H_k = \max_{j \in M} H_j\}.$$

Here establishment of set D^0 allows us to deal with the situation when some of values H_i may be the same.

The risk (9) means the probability not to assign to hypothetical distribution (4) those samples of $\{X^{(i)}\}_{i \in M}$ that are closer to (4) in terms of the distribution similarity expressed in values (6).

If all values $\{H_i\}_{i \in M}$ are distinguished then the risk (9) is simplified:

$$r = r(d(X^{(1)}, \ldots, X^{(m)})) = P\{d(X^{(1)}, \ldots, X^{(m)}) \neq d^0\}; \tag{10}$$

$$d^0 = \arg\max_{i \in M} H_i.$$

3 The asymptotical investigation of the risk in the case of two samples of the same size. The Fisher model.

Now let us assume the situation when there are only two ($m = 2$) samples $X^{(1)} = \{x_t^{(1)}\}_{t=1}^n$, $X^{(2)} = \{x_t^{(2)}\}_{t=1}^n$ of the same size ($n_1 = n_2 = n$) given for assignment to the hypothetical distribution (2). Then it becomes possible to rewrite DR (4) in the form:

$$d(X^{(1)}, X^{(2)}) = \begin{cases} 1, & if \quad \overline{\xi}_n(X^{(1)}, X^{(2)}) \leq 0; \\ 2, & if \quad \overline{\xi}_n(X^{(1)}, X^{(2)}) > 0, \end{cases} \tag{11}$$

where

$$\overline{\xi}_n(X^{(1)}, X^{(2)}) = \frac{1}{n} \sum_{t=1}^{n} \ln \frac{p(x_t^{(2)})}{p(x_t^{(1)})} \tag{12}$$

and $p(\cdot)$ is the hypothetical probability density from (2).

Also the risk r (9), (10) of the decision rule (11), (12) takes form:

$$r = \begin{cases} P\{\overline{\xi}_n(X^{(1)}, X^{(2)}) \leq 0\}, & if \quad H_1 < H_2; \\ 1 - P\{\overline{\xi}_n(X^{(1)}, X^{(2)}) \leq 0\}, & if \quad H_1 > H_2; \\ 0, & if \quad H_1 = H_2, \end{cases} \tag{13}$$

where H_1, H_2 are values from (6).

Theorem. *Let us consider the assignment problem of two samples ($m = 2$) of the same size ($n_1 = n_2 = n$) and let the following conditions be true:*

$$G_i = \int_{R^N} (\ln(p(x)))^2 p_i(x) dx < +\infty, \quad G_i - H_i^2 \neq 0, \quad i = 1, 2, \tag{14}$$

where $p_1(\cdot)$, $p_2(\cdot)$ and $p(\cdot)$ are densities from (1), (2).

Then the risk (13) can be calculated asymptotically (assuming $H_1 \neq H_2$):

$$\frac{r}{\tilde{r}} \to 1, \quad n \to +\infty; \quad \tilde{r} = \Phi\left(-\sqrt{n}\frac{|H_1 - H_2|}{\sqrt{G_1 + G_2 - (H_1^2 + H_2^2)}}\right), \tag{15}$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{w^2}{2}\right) dw, \quad z \in R,$$

is the standard Gaussian distribution function.

For further results let us assume that all densities $p_1(\cdot)$, $p_2(\cdot)$ and $p(\cdot)$ are multivariate Gaussian with the same covariance matrix. Such assumption is often used in various applications and it is known as the Fisher model [1, 3, 4]:

$$p_i(x) = n_N(x|\mu_i, \Sigma), \quad i = 1, 2; \tag{16}$$

$$p(x) = n_N(x|\mu, \Sigma);$$

$$n_N(x|\mu, \Sigma) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)\right), \quad x \in R^N,$$

where

$$\mu_i = \int_{R^N} x\, p_i(x) dx, \quad i = 1, 2; \quad \mu = \int_{R^N} x\, p(x) dx$$

are appropriate mathematical mean N-vectors and

$$\Sigma = \mathbf{E}\{(x - \mu_i)(x - \mu_i)' | d^o = i\}, \quad i \in S,$$

is the common non-singular covariance $(N \times N)$-matrix.

Under the Fisher model (16) the asymptotical risk $\tilde{r}$ (15) takes the form:

$$\tilde{r} = \Phi\left(-\sqrt{n}\frac{|\rho^2(\mu, \mu_1) - \rho^2(\mu, \mu_2)|}{2\sqrt{N + \rho^2(\mu, \mu_1) + \rho^2(\mu, \mu_2)}}\right), \tag{17}$$

where $\rho(\mu, \mu_i) = \sqrt{(\mu - \mu_i)'\Sigma^{-1}(\mu - \mu_i)}$ is the Mahalanobis distance [1, 3, 4] between μ and μ_i $(i = 1, 2)$.

References

[1] Kharin Yu.S., Zuev N.M., Zhuk E.E. (2011). *Probability Theory, Mathematical and Applied Statistics*. BSU, Minsk.

[2] Borovkov A.A. (1984). *Mathematical Statistics*. Nauka, Moskow.

[3] Aivazyan S.A., Buchstaber V.M., Yenyukov I.S., Meshalkin L.D. (1989). *Applied Statistics: Classification and Dimensionality Reduction*. Finansy i Statistika, Moskow.

[4] Zhuk E.E. (2013). Assignment of multivariate samples to the fixed classes by the maximum likelihood method and its risk. *Computer Data Analysis and Modeling: Theoretical and Applied Stochastics : Proc. of the Tenth Intern. Conf.* Vol. **1**, pp. 185-188.