ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
21,2 (1991), pp. 175–178

S. TRYBUŁA (Wrocław)

MINIMAX PREDICTION OF THE DIFFERENCE OF SAMPLE DISTRIBUTION FUNCTIONS

In the paper a minimax predictor of the difference of sample distribution functions is determined for the loss function (1).

Suppose that the random variables X and U are distributed according to the unknown distribution F, and the random variables Y and V are distributed according to the unknown distribution G. Let $\overline{X} = (X_1, \ldots, X_m)$, $\overline{U} = (U_1, \ldots, U_n)$ and $\overline{Y} = (Y_1, \ldots, Y_m)$, $\overline{V} = (V_1, \ldots, V_n)$ be independent random samples from F and G, respectively. Set

$$\check{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{U_i}(t), \quad \check{G}(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{V_i}(t),$$

where, for a random variable Z,

$$\delta_Z(t) = \begin{cases} 1 & \text{if } Z \le t, \\ 0 & \text{if } Z > t. \end{cases}$$

Let $\varphi(t) = \varphi(t; \overline{X}, \overline{Y})$ be a predictor of $\check{F}(t) - \check{G}(t)$. We suppose that the loss function associated with the predictor $\varphi(t)$ is

$$L(\check{F}, \check{G}, \varphi) = \int_{-\infty}^{\infty} (\varphi(t) - \check{F}(t) + \check{G}(t))^2 \, w(dt), \tag{1}$$

where w is a non-zero finite measure on $(\mathbf{R}, \mathcal{B})$, $\mathcal{B}$ being the σ-field of Borel subsets of $\mathbf{R} = (-\infty, \infty)$.

We solve the problem of determining a minimax predictor of $\check{F}(t) - \check{G}(t)$ for this loss function.

Let us study a predictor of the form

$$\varphi(t) = a(\widehat{F}(t) - \widehat{G}(t)),$$

1991 *Mathematics Subject Classification*: Primary 62F15.
Key words and phrases: minimax prediction, cumulative distribution function, sample distribution function.

where

$$\hat{F}(t) = \frac{1}{m}\sum_{i=1}^{m} \delta_{X_i}(t), \quad \hat{G}(t) = \frac{1}{m}\sum_{i=1}^{m} \delta_{Y_i}(t).$$

The risk function for this predictor is

$$\begin{aligned} R(F,G,\varphi) &= E[L(\check{F},\check{G},\varphi(t;\overline{X},\overline{Y}))] \\ &= \int_{-\infty}^{\infty} E[(a(\hat{F}(t)-\hat{G}(t)) - (\check{F}(t)-\check{G}(t)))^2]\, w(dt) \\ &= \int_{-\infty}^{\infty} \left\{ \left(\frac{a^2}{m}+\frac{1}{n}\right)[F(t)(1-F(t)) + G(t)(1-G(t))] \right. \\ &\qquad \left. + (1-a)^2[F(t)-G(t)]^2 \right\} w(dt). \end{aligned}$$

Let

$$\frac{a^2}{m} + \frac{1}{n} = 2(1-a)^2,$$

which is satisfied if

$$a = \frac{2-\frac{1}{n}}{2+\sqrt{\frac{2}{m}+\frac{2}{n}-\frac{1}{mn}}}. \tag{2}$$

For this a the risk is

$$R(F,G,\varphi) = (1-a)^2 \int_{-\infty}^{\infty} [2(F(t)+G(t)) - (F(t)+G(t))^2]\, w(dt).$$

For fixed t the expression in square brackets attains its maximum when $F(t)+G(t)=1$. Then for

$$\varphi_0(t) = \frac{\left(2-\frac{1}{n}\right)(\hat{F}(t)-\hat{G}(t))}{2+\sqrt{\frac{2}{m}+\frac{2}{n}-\frac{1}{mn}}} \tag{3}$$

we obtain

$$R(F,G,\varphi_0) \le (1-a)^2 \int_{-\infty}^{\infty} w(dt) \stackrel{\mathrm{df}}{=} c. \tag{4}$$

We shall prove that the predictor $\varphi_0(t)$ is minimax.

The considered problem of determining a minimax predictor of $\check{F}(t) - \check{G}(t)$ can be viewed as a problem of finding the optimal strategy in a game against nature. The nature chooses cumulative distribution functions $F(t)$

and $G(t)$, the statistician chooses a predictor $\varphi(t)$ and the payoff function is given by (1). Let us define a sequence τ_k of mixed strategies of nature which will be used in the proof of the optimality of the strategy $\varphi_0(t)$.

Choose the parameter p according to the density

$$(5) \qquad g(p) = \begin{cases} \dfrac{1}{B(\alpha,\alpha)}[p(1-p)]^{\alpha-1} & \text{if } 0 < p < 1, \\ 0 & \text{otherwise}, \end{cases}$$

and then, for given p, choose the distributions $F(t)$ and $G(t)$ of the form

$$(6) \quad F(t) = \begin{cases} 0 & \text{if } t < -k, \\ p & \text{if } -k \le t < k, \\ 1 & \text{if } t \ge k, \end{cases} \qquad G(t) = \begin{cases} 0 & \text{if } t < -k, \\ 1-p & \text{if } -k \le t < k, \\ 1 & \text{if } t \ge k. \end{cases}$$

For any predictor $\varphi(t)$ we have

$$R(F,G,\varphi) = \int_{-\infty}^{\infty} \left[E(\varphi(t) - F(t) + G(t))^2 + \frac{F(t)(1-G(t))}{n} + \frac{G(t)(1-G(t))}{n} \right] w(dt).$$

Notice that the second and the third terms in square brackets do not depend on $\varphi(t)$.

Let $F(t)$ and $G(t)$ be given by (6), where the parameter p has the distribution (5). For the strategy τ_k the expected risk is

$$r(\tau_k, \varphi) = \int_{-\infty}^{\infty} E_{\tau_k}[E(\varphi(t) - F(t) + G(t))^2]\, w(dt) + r_0(\tau_k),$$

where $r_0(\tau_k)$ does not depend on φ and $E_{\tau_k}(\cdot)$ is the expectation with respect to the density $g(p)$ in the strategy τ_k. Thus, in order to minimize the expected risk $r(\tau_k, \varphi)$, it is sufficient to minimize $E_{\tau_k}[E(\varphi(t) - F(t) + G(t))^2]$ for any fixed t. This leads to the Bayes predictor with respect to τ_k

$$\varphi_{\tau_k}(t) = \begin{cases} 0 & \text{if } t < -k, \\ m\dfrac{\widehat{F}(t) - \widehat{G}(t)}{m+\alpha} & \text{if } -k \le t < k, \\ 0 & \text{if } t > k. \end{cases}$$

Assume

$$\alpha = \frac{2mn}{2n-1}\left(\frac{1}{n} + \sqrt{\frac{2}{m} + \frac{2}{n} - \frac{1}{mn}}\right).$$

In this case $\varphi_{\tau_k}(t) = \varphi_0(t)$ if $-k \le t < k$, and since under τ_k we have $F(t) + G(t) = 1$ for $-k \le t < k$, the Bayes risk $r(\tau_k, \varphi_{\tau_k})$ is

$$r(\tau_k, \varphi_{\tau_k}) = (1-a)^2 \int_{-\infty}^{\infty} I_{(-k,k)}(t)\, w(dt),$$

where $I_A(t)$ is the characteristic function of the set A and a is given by (2) (compare with (4)).

From the above it follows that

$$\lim_{R\to\infty} r(\tau_k, \varphi_{\tau_k}) = c. \tag{7}$$

From (4) *and* (7) *it follows that* $\varphi_0(t)$ *given by* (3) *is a minimax predictor of* $\check{F}(t) - \check{G}(t)$.

If the measure w is concentrated at one point, say t_0, then the problem reduces to that of determining a minimax predictor of the random variable $\check{F}(t_0) - \check{G}(t_0)$, which is, after multiplication by n, a difference of binomial random variables.

For problems of estimation of a cumulative distribution function see [1], [2], [4], [5]. Minimax estimators of a cumulative distribution function for four loss functions of type (1) were found by Phadia in [3].

References

[1] O. P. Aggarwal, *Some minimax invariant procedures for estimating a cumulative distribution function*, Ann. Math. Statist. 26 (1955), 450–462.

[2] A. Dvoretzky, J. Kiefer and J. Wolfowitz, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator*, ibid. 27 (1956), 642–669.

[3] E. G. Phadia, *Minimax estimation of a cumulative distributions function*, Ann. Statist. 1 (1973), 1149–1157.

[4] R. R. Read, *The asymptotic inadmissibility of the sample distribution function*, Ann. Math. Statist. 42 (1972), 89–95.

[5] S. Trybuła, *Estimation of the difference of cumulative distribution functions*, Bull. Polish Acad. Sci. Math. 32 (1984), 243–246.

STANISŁAW TRYBUŁA
INSTITUTE OF MATHEMATICS
TECHNICAL UNIVERSITY OF WROCŁAW
WYBRZEŻE WYSPIAŃSKIEGO 27
50-370 WROCŁAW, POLAND

Received on 11.7.1990