UDC 519.234.22

On Robust Sequential Parameters Estimating

I. I. Tsitovich[*†]

[*] *Institute for Information Transmission Problems*
(Kharkevich Institute) RAS,
Bolshoy Karetny per. 19, build.1, Moscow, 127051, Russia
[†] *Faculty of Computer Science,*
National Research University Higher School of Economics,
Myasnitskaya ul., 20, Moscow, 101000, Russia

Abstract. We study the problem of parameters estimating if there is a slight deviation between the parametric model and real distributions. The estimator is based on suboptimal testing of builded by a special way nonparametric hypotheses. It is proposed a natural for this problem risk function. We found that the risk function has an exponential decrease to the mean number of observations. Numerical results of a comparative analysis our risk function behaviour for proposed estimator and some another estimators are outlined.

Keywords: estimating, robustness, sequential analysis, suboptimality.

1. Introduction

Robust estimation of a statistical model parameters is one of important problems in the statistic. The main problem consists in a rapidly decreasing power of a robust estimator under deviations from the pure parametric model. Our approach is the more strong, additionally we investigate a problem to construct a guaranteeing decision.

One of popular method for a robust estimator construction bases on influence functions [1]. But it is known that a power of a statistical decision depends on a distribution tail. In [2]– [4] there are investigated a problem of robust discriminating of hypotheses and an influence of a tail decreasing on a test power are indicated. Based on modification of the sequential probability ratio test, there are obtained the suboptimal sequential test. It is so that its power converges to the power of asymptotically optimal sequential test when the neighborhood size of the hypothesis converges to 0. In this paper, we apply this method for a robust estimation.

2. Setting of the problem

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and $x_1, x_2, \ldots$ be independent identical distributed random variables with values in a subset $X \subset \mathbf{R}$ with the probability distribution P from a set $\mathcal{P}$. Let be some nondegenerate measure μ on X (with Borel σ-algebra on X) such that any probability distribution $\mathsf{P} = \mathsf{P}_f$ from the set $\mathcal{P}$ has its density function $f(x)$ with respect to μ.

We suppose that $\mathcal{P}$ has the following structure.

Let $z(\mathsf{P},\mathsf{Q},x) = \log\frac{p(x)}{q(x)}$ and $I(\mathsf{P},\mathsf{Q}) = \mathsf{E}_{\mathsf{P}} z(\mathsf{P},\mathsf{Q},x)$ be the relative entropy (Kullback–Leibler divergence) with usual conventions (logarithms are to the base e, $0\log 0 = 0$ etc).

C1. It is exists a metric d on $\mathcal{P}$ such that I is uniformly continuous on $\mathcal{P}$ under d.

Let $\mathcal{P}_0$ be a parametric set of densities $f(\theta, x)$ with respect to μ, $\theta \in \Theta$, where Θ is a compact in R^m and $\mathcal{P}_0 \subset \mathcal{P}$, and the set $\mathcal{P}_0$ be continuous under the metric d. The distribution with the density $f(\theta, x)$ is denoted as P_θ.

Let us define for $\mathsf{P} \in \mathcal{P}_0$ a neighborhood $\mathcal{O}_\delta(\mathsf{P}) = \left\{\widetilde{\mathsf{P}}\right\}$, δ is a positive number, as the subset in $\mathcal{P}$ all distributions $\widetilde{\mathsf{P}}$ where $d(\mathsf{P},\widetilde{\mathsf{P}}) < \delta$. Let

$$\mathcal{P}_\delta = \bigcup_{\theta\in\Theta} \mathcal{O}_\delta(\mathsf{P}_\theta).$$

Therefore, the set $\mathcal{P}_\delta$ is the δ-neighborhood of the parametric family $\mathcal{P}_0$ in $\mathcal{P}$.

We suppose that $\mathcal{P} = \mathcal{P}_\delta$ for certain $\delta > 0$.

The neighborhoods $\mathcal{O}_\delta(\mathsf{P}_\theta)$, $\theta \in \Theta$, generate open sets $\mathcal{O}'_\delta(\mathsf{P}_\theta) = \mathcal{O}_\delta(\mathsf{P}_\theta)\bigcap\mathcal{P}_0$ in $\mathcal{P}_0$ and $\mathcal{O}'_\delta(\mathsf{P}_\theta)$, $\theta \in \Theta$, give an open cover of $\mathcal{P}_0$ and this cover has finite subcovers. Any subcover is described by the set of its neighborhoods centers $\{\theta_1, \ldots, \theta_m\}$.

Let us fix a subcover and denote its neighborhoods centers set as $\{\theta_1^0, \ldots, \theta_{m_0}^0\}$. The subcover has two characteristics: the accuracy of the parameter estimating

$$\Delta^0 = \max_i \min_{j\in A(i)} |\theta_i^0 - \theta_j^0|$$

and the information distance

$$I_i = \inf_{\mathsf{P}\in\mathcal{O}_\delta(\mathsf{P}_{\theta_i^0})} \inf_{\mathsf{Q}\in\bigcup_{j\in A(i)}\mathcal{O}_\delta(\mathsf{P}^0_{\theta_j})} I(\mathsf{P},\mathsf{Q}),$$

where $A(i)$ is the alternative set of parameters for θ_i^0 and defines as

$$A(i) = \{j:\ \mathcal{O}_\delta(\mathsf{P}_{\theta_i^0})\bigcap\mathcal{O}_\delta(\mathsf{P}_{\theta_j^0}) = \emptyset\}.$$

We have two contradict requirements for the subcover: Δ^0 needs be as large as possible for maximizing I_i and it needs be as small as possible for maximizing accuracy of the parameters estimating.

All $\theta' \in \mathcal{O}'_\delta(\mathsf{P}_\theta)$ are undistinguished by the accuracy of the statistical model δ and, therefore, a risk function of the parameter estimation $R(\theta,\theta')$ needs be 0 for $\theta' \in \mathcal{O}'_\delta(\mathsf{P}_\theta)$. By this reason we define the lost function for an estimator $\hat\theta$ as

$$R(\hat\theta) = \sup_{\mathsf{P}\in\mathcal{P}_\delta} \mathbf{P}_{\mathsf{P}}(|\theta^0_{i(\mathsf{P})} - \hat\theta| > \Delta). \quad (1)$$

This lost function means that we find guarantee decisions only.

3. Main result

We use the following regularity conditions:
C2. There is $c > 0$ such that $\mathbf{E}_{\mathsf{P}} \left(z(\mathsf{P}, \mathsf{Q}, x_i)\right)^2 < c$ for all $\mathsf{P} \in \mathcal{P}, \mathsf{Q} \in \mathcal{P}$.
C3. There exist $t > 0$ and $f > 0$ such that for all $\mathsf{P} \in \mathcal{P}$

$$\mathbf{E}_{\mathsf{P}} \left(\sup_{\mathsf{Q} \in \mathcal{P}} \exp(-tz(\mathsf{P}, \mathsf{Q}, x_i)) \right) \leq f.$$

C4. $z(\mathsf{P}, \mathsf{Q}, x)$ is differentiable w.r.t. x and

$$D = \int_X z_1(x) \left(a(x) b(x)\right)^{1/2} dx < \infty,$$

where

$$z_1(x) = \sup_{\mathsf{Q} \in \mathcal{P}} \left| \frac{\partial z(\mathsf{P}, \mathsf{Q}, x)}{\partial x} \right|,$$

$$\sup_{\mathsf{P} \in \mathcal{P}} \int_{-\infty}^{x} p(t) \mu(dt) \leq a(x), \quad \sup_{\mathsf{P} \in \mathcal{P}} \int_{x}^{\infty} p(t) \mu(dt) \leq b(x).$$

Let us introduce $L_k(\mathsf{P}, \mathsf{Q}) = \sum_{i=1}^{k} z(\mathsf{P}, \mathsf{Q}, x_i)$.

We perform the following estimator $\tilde{\theta}$. We stop observations at the first moment M such that

$$\max_i \inf_{\mathsf{Q} \in A(i)} L_M(\mathsf{P}_{\theta_i^0}, \mathsf{Q}) > -\log \beta \tag{2}$$

and accept as an estimation $\tilde{\theta} = \theta_r^0$ if (2) holds and r is the value of argmax_i.

Let

$$k(\mathsf{P}) = \inf_{\mathsf{Q} \in \bigcup_{j \in A(i)} \mathcal{O}_\delta(\mathsf{P}_{\theta_j})} I(\mathsf{P}, \mathsf{Q}), e = \max_i \sup_{\mathsf{Q} \in \mathcal{O}_\delta(\mathsf{P}_{\theta_j^0})} I(\mathsf{Q}, \mathsf{P}_{\theta_j^0}),$$

where i such that $\mathsf{P} \in \mathcal{O}_\delta(\mathsf{P}_{\theta_i})$. If there are several i such that $\mathsf{P} \in \mathcal{O}_\delta(\mathsf{P}_{\theta_i})$ then we take i such that the respective value $k(\mathsf{P})$ is maximal.

Based on results of [5, 6] we get the following results

Theorem 1. *If $\mathsf{P} \in \mathcal{P}_\delta$ is such that $k(\mathsf{P}) > e$ then under the conditions C1–C4*

$$\mathbf{E}_{\mathsf{P}} M \leq \frac{|\log \beta|}{k(\mathsf{P}) - e} + K \sqrt{|\log \beta|}$$

with the same constant K for all $\beta > 0$ and $\mathsf{P} \in \mathcal{P}_\delta$.

Theorem 2. *Under the condition C1*

$$\sup_{\mathsf{P}\in\mathcal{P}} \mathbf{P}_{\mathsf{P}}(|\theta_i - \tilde{\theta}| > \Delta) \le m_0\beta.$$

From this result follows that the risk function (1) can be estimating.

4. Numerical results

For numerical illustration of the general theory we propose the simplest example of estimating the mean of the normal distribution, i.e. $f(\theta, x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right)$, $\Theta = [\theta_l,\, \theta_r], \theta_l < 0 < \theta_r$. This distribution is mixed with another distribution with a density $g(x)$ with respect to the Lebesgue measure, $X = \mathsf{R}$, by the formula $h(x) = (1-\varepsilon)f(\theta, x) + \varepsilon g(x)$, where ε, $0 \le \varepsilon < 1$, is a parameter of this mixture, and is denoted as H_ε.

Let θ_c be the estimation when the estimator $\bar{x}$ is based on censored data, θ_w be the estimation when the estimator $\bar{x}$ is based on winsored data when appropriate levels are A_l (downer) and A_r (upper). This means that we reduce $X = \mathsf{R}$ to the segment $[A_l; A_r]$, the distribution P_θ has the density $f(\theta, x)$ for $x \in (A_l; A_r)$ and atoms $\Phi(A_l - \theta)$ and $1 - \Phi(A_r - \theta)$ at the points A_l and A_r respectively, the the distribution with the density $g(x)$ has the atoms $G_l = \int_{-\infty}^{A_l} g(x)dx$ and $G_r = \int_{A_r}^{\infty} g(x)dx$. The neighborhood $\mathcal{O}_\delta(\mathsf{P}_\theta)$ is

$$\mathcal{O}_\delta(\mathsf{P}_\theta) = \Big\{ h(x) \mid \forall\, x \in (A_l; A_r), \quad |h(x) - f(\theta, x)| \le \delta f(\theta, x) \Big\}.$$

We outline one example of numerical investigations with the following parameters $A_l = -2$, $A_r = 2$, $\theta_l = -1$, $\theta_r = 1$, $\varepsilon = 0.1$, $g(x) = \frac{1}{\pi}\frac{1}{1+x^2}$, $\Delta = 0.2$, true value of the parameter θ is 0.

Table 1

Results of estimating when $g(x) = \frac{1}{\pi}\frac{1}{1+x^2}$

L	$R(\tilde{\theta})$	$R(\theta_w)$	$\mathrm{R}(\theta_c)$	$\bar{M}$
2	0.1065	0.2652	0.1886	23.93
4	0.0223	0.1248	0.0625	46.73
6	0.0045	0.0607	0.0246	70.11
8	0.0007	0.0307	0.0091	93.88
10	0.0002	0.0152	0.0042	118.46

If $\varepsilon = 0$ and we test the hypothesis $\theta = 0$ under the alternative $\theta = \Delta$ then $k(\mathsf{H}_0) = \frac{\Delta^2}{2}$. It is natural to use as a measure of effectiveness of an estimator $\hat{\theta}$

$$\mathcal{E}(\hat{\theta}, \mathsf{P}) = \lim_{\beta \to 0} \frac{-2\log(R(\hat{\theta}, \mathsf{P}))}{\Delta^2 \bar{M}}.$$

Then $\mathcal{E}(\bar{x}, \mathsf{H}_0) = 1$, where $\bar{x}$ is the standard estimation of the mean, and it is followed from Table 1 that $\mathcal{E}(\tilde{\theta}, \mathsf{H}_{0.1}) \approx 0.8$, $\mathcal{E}(\theta_w, \mathsf{H}_{0.1}) \approx 0.39$, $\mathcal{E}(\theta_c, \mathsf{H}_{0.1}) \approx 0.52$.

5. Conclusions

We propose the setting of the problem of sequential robust estimating of an unknown parameters with a guaranteeing decision and the risk function of an estimation.

For this setting we construct the estimator with near to optimal properties for some statistical models.

It is found that, in general, the rate of the risk function decreasing is an exponential under the mean number of observations.

For constructing an effective estimator, it is necessary to find a cover of the parametric space in such a way that optimizing the information distance to the alternative set of parameters for a given level of an estimating accuracy Δ.

References

1. *Huber P. J.* Robust Statistics. — Wiley, 2004.
2. *Tsitovich F. I.* Properties of suboptimal sequential decisions for testing nonparametric hypotheses on distributions with exponentional decriasing tails // Information Processes. — 2010. — Vol. 10, no 2. — P. 181–196. (in Russian)
3. *Tsitovich F.* Supoptimal Multistage Nonparametric Hypotheses Test // Pliska. Studia mathematica Bulgaria. — 2009. — Vol. 19. — P. 269–282.
4. *Tsitovich F., Tsitovich I.* Sample space reducing for statistical decision effectiveness increasing // 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). — IEEE, 2014. — ISBN: 978-1-4799-5291-5. — P. 501–506.
5. *Malyutov, M.B., Tsitovich, I.I.* Second order optimal sequential tests // Proc. Internat. Workshop Optimum Design 2000, Cardiff, UK, April 2000. — Kluwer, Netherlands, 2000. — P. 67–78.
6. *Malyutov M. B., Tsitovich I. I.* Second Order Optimal Sequential Model Choice and Change-point Detection // Information Processes. — 2010. — Vol. 10, no 3. — P. 275–291. (in Russian)