Robustification of Sequential Statistical Decision Rules for Stochastic Data Flows Analysis

Alexey Kharin
Department of Probability Theory and Mathematical Statistics
Belarusian State University
Minsk, Belarus
KharinAY@bsu.by

Ton That Tu
Division of Applied Mathematics
University of Danang
Danang, Vietnam
tthattu@gmail.com

Dai Yukun
Faculty of Applied Mathematics and Computer Sciences
Belarusian State University
Minsk, Belarus
daiyukun0905@gmail.com

Wang Yumin
Faculty of Applied Mathematics and Computer Sciences
Belarusian State University
Minsk, Belarus
wangyumin1994@gmail.com

Abstract. **In data analysis the issues of statistical decision making on parameters of observed stochastic data flows are important. To solve the relevant problems, here sequential statistical decision rules are used. The sequential statistical decision rules traditionally used loose their performance optimality in situations that are common in practice, when the hypothetical model is distorted. Here the robustified sequential decision rules are constructed for three models of observation flows: independent homogeneous observations; observations forming a time series with a trend; dependent observations forming a homogeneous Markov chain.**

Index Terms: **sequential decision rule, time series with trend, homogeneous Markov chain, distortion, robustness**

I. Introduction

In applications, a problem of stochastic data flows analysis is often important [1], aiming decisions on one of two possible modes of a system that generates such a flow. The mode corresponds to a parameter value (or a value of parameters vector) that defines the probabilistic properties of the observed data.

To construct efficient decision rules, one of possible approaches is the sequential statistical analysis [2]. It exploits the key concept concerning the number of observations needed to make a decision with given small levels of error probabilities. It is not fixed a priori, and is defined through the observation process on the basis of the observed random values, so it is random itself [3]. The number of observations is tailored for each situation with the observed data flow, and this feature makes effective the resulting decision rule. There are two admissible decisions after each observation received: to stop the process and to decide in favor of one of two defined hypotheses, or to collect the next observation, as the requested accuracy is not reached at the moment. Although theoretical analysis of sequential decision rules is not trivial, it is widely used in medicine, finance, quality control, and other fields, where the cost of each observation is not ignorable.

Here we use the approach developed in [4], [7], to construct robustified sequential decision rules, i.e. the rules that are robust [4], [5] under distortions of the hypothetical model of data [6]. In other words, their performance characteristics are essentially less influenced by the distortions as compared to the sequential decision rules that are traditionally used.

II. Sequential Decision Rule for the Flow of Independent Homogeneous Observations

Let a data flow of independent random vectors $x_1, x_2, \dots$ be observed with a probability distribution P_θ that has the probability density function $p_\theta(x)$, $x \in U \subseteq R^N$, where $\theta \in \Theta = \{0, 1\}$ is a parameter value which is not observed.

There are two hypotheses corresponding to two modes of the system, in terms of the parameter value:

$$H_0 : \ \theta = 0, \ H_1 : \ \theta = 1.$$

A sequential decision rule is defined as a pair of components: stopping moment rule, and acceptance (terminal decision) rule. Consider a family of sequential decision rules $\delta_\lambda = (\tau_\lambda, d_\lambda)$ based on function $\lambda(\cdot)$: $U \to R$, where

$$\tau_\lambda = \inf\{n : \ \Lambda_n \notin (C_-, C_+)\}$$

is the stopping moment rule (the result depends on $x_1, \dots, x_n$ and that is why it is random), and

$$d_\lambda = 1_{[C_+, +\infty)}(\Lambda_n)$$

is the terminal decision rule in favor of the hypothesis H_i, if $d_\lambda = i$, $i \in \{0, 1\}$.

The test statistic is

$$\Lambda_n = \Lambda_n(x_1, \dots, x_n) = \sum_{t=1}^{n} \lambda(x_t), \ n \in N = \{1, 2, \dots\}.$$

The research is supported by the State research program "Digital and space technologies, human and state safety".

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

The parameters $C_-, C_+ \in R$, $C_- < C_+$ are called thresholds and calculated according to:

$$C_- = \log \frac{\beta}{1-\alpha},\ C_+ = \log \frac{1-\beta}{\alpha},$$

where α, β are values given by a user for admissible levels of error type I (H_0 is true, but declined) and II (H_1 is true, but declined) probabilities.

For the traditionally used sequential probability ratio test

$$\lambda(u) = \log \frac{p_1(u)}{p_0(u)},\ u \in U.$$

Let the described above hypothetical model be distorted:

$$\bar{P}_k(x) = (1-\epsilon_k)P_k(x) + \epsilon_k \tilde{P}_k(x),\ x \in U,\ k \in \{0,1\}$$

be the factual data flow observations probability distribution, representing a mixture of the hypothetical probability distribution P_k and of the contaminating probability distribution $\tilde{P}_k$, where $\epsilon_k \in [0, \frac{1}{2})$ is the probability of contamination (contamination level).

To construct the robust sequential decision rule, a family of modified sequential tests $\delta_g = (\tau_g, d_g)$ is developed:

$$\tau_g = \inf\{n :\ \sum_{t=1}^{n} g(\lambda_W(x_t)) \notin (C_-, C_+)\},$$

$$d_g = 1_{[C_+,+\infty)}(\sum_{t=1}^{n} g(\lambda_W(x_t))),$$

where

$$g(z) = g_- 1_{(-\infty,g_-)}(z) + z \cdot 1_{[g_-,g_+]}(z) + g_+ 1_{(g_+,+\infty)}(z),$$
$$z \in R.$$

Here $g_-, g_+ \in R$ are extra parameters of the developed sequential decision rules. Using this paprameters, the robustified sequential decision rules are constructed with minimax criterion w.r.t. the risk function.

III. Inhomogeneous Data Flows Forming Time Series with a Trend

Consider the model of stochastic data flow, where inhomogeneous independent observations forming a time series with a trend [8].

Let $x_1, x_2, \ldots$ be observations of a time series with a trend:

$$x_t = \theta^T \psi(t) + \xi_t,\ t \geq 1,$$

where $\psi(t) = (\psi_1(t), \ldots, \psi_l(t))$, $t \geq 1$, are the vectors of the trend basic functions, $\theta = (\theta_1, \ldots, \theta_l)^T \in \mathbf{R}^l$ is a vector of coefficients, their values are not known along with the observation process, $\{\xi_t,\ t \geq 1\}$ is a sequence of independent identically distributed random variables from $\mathcal{N}_1(0, \sigma^2)$.

To give more flexibility in the decision making, the case of M simple hypotheses is considered w.r.t. the vector θ. The following two sequential test were analyzed.

A. M-ary sequential probability ratio test. It uses the posterior probabilities of the hypotheses. The stopping time N_a and the final decision d_a for this test are defined by the equations:

$$N_a = \inf\Big\{ n \geq 1 :\ \exists m \in \{1, \ldots, M\},$$

$$P\{\mathcal{H}_m \mid x_1, \ldots, x_n\} > \frac{1}{1+A_m} \Big\},$$

$$d_a = \arg \max_{1 \leq m \leq M} P\{\mathcal{H}_m \mid x_1, \ldots, x_{N_a}\},$$

where $A_m \in (0, 1]$ are some specified constants, $m \in \{1, \ldots, M\}$, $d_a = m$ means that the decision in favor of the hypothesis H_m is made.

B. Matrix sequential probability ratio test. Denote

$$\Lambda_n(i,j) = \ln \left(\prod_{t=1}^{n} \frac{n_1(x_t; (\theta_i)^T \psi(t), \sigma^2)}{n_1(x_t; (\theta_j)^T \psi(t), \sigma^2)} \right);$$

$$\tau_i = \inf\{n \in \mathbf{N} :\ \Lambda_n(i,j) > b_{ij},$$

$$\forall j \in \{1, \ldots, M\} \setminus \{j\}\},\ i \in 1, \ldots, M,$$

where $B = (b_{ij})$, $i, j \in \{1, \ldots, M\}$, is the matrix of the test thresholds (using them, the error probabilities of the test are controlled by the user f the decision rule). For this test the stopping time N_b and the final decision d_b are defined as follows:

$$N_b = \min\{\tau_i :\ i \in \{1, \ldots, M\}\},\ d_b = \arg \min_{i \in \{1,\ldots,M\}} \tau_i.$$

For the two sequential tests defined above, the termination with probability 1 property and the finiteness of all moments of the random stopping time are proved under a condition reasonable and affordable for practice. For the M-ary sequential probability ratio test, upper bounds for the error probabilities are derived.

A robustified version of the matrix sequential probability ratio test based on change limitation for the test statistics is constructed and its properties are analyzed via numerical experiments.

IV. Dependent Data Flow Forming a Homogeneous Markov Chain

Consider here the situation, where observations are dependent and forming a homogeneous Markov chain [4].

Let the data flow be dependent observations forming a homogeneous Markov chain $x_1, x_2, \ldots$, with possible values in the set $V = \{0, 1, \ldots, M-1\}$. Denote the vector of initial states probabilities by $\pi = (\pi_i)$, $i \in V$, and the one-step transition probabilities matrix by $P = (p_{ij})$, $i, j \in V$, that are: $\mathrm{P}\{x_1 = i\} = \pi_i$, $\mathrm{P}\{x_n = j \mid x_{n-1} = i\} = p_{ij}$, $i, j \in V$, $n > 1$.

There are two hypotheses concerning the Markov chain parameters introduced above: $\mathcal{H}_0$: $\pi = \pi^{(0)}$, $P = P^{(0)}$ with the alternative $\mathcal{H}_1$: $\pi = \pi^{(1)}$, $P = P^{(1)}$, where $\pi^{(0)}$, $\pi^{(1)}$ are the given values of the initial states probabilities vector,

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

$P^{(0)} \neq P^{(1)}$ are the one-step transition probabilities matrices for correspondent hypotheses. Denote also:

$$\lambda_1 = \ln \frac{P_1\{x_1\}}{P_0\{x_1\}},\ \lambda_k = \ln \frac{P_1\{x_k \mid x_{k-1}\}}{P_0\{x_k \mid x_{k-1}\}},\ k > 1,$$

$$\Lambda_n = \sum_{k=1}^{n} \lambda_k,\ n \in \mathbb{N},$$

where $P_s\{x_1\}$ is the probability to observe the value x_1, $P_s\{x_k \mid x_{k-1}\}$ is the probability to observe x_k at the moment k provided at the moment $k-1$ the value x_{k-1} was observed, if hypothesis $\mathcal{H}_s$, is true $s \in \{0, 1\}$.

As it was done above, construct the sequential decision rule to decide in favor of $\mathcal{H}_0$ or $\mathcal{H}_1$. According to this decision rule, with given thresholds values $C_-, C_+ \in \mathbb{R}$, $C_- < 0$, $C_+ > 0$, hypothesis $\mathcal{H}_0$ is accepted on the basis of n observations, if $\Lambda_n \leq C_-$. Hypothesis $\mathcal{H}_1$ is accepted, if $\Lambda_n \geq C_+$, otherwise the observation process is not stopped, and $(n+1)$-th observation is requested.

Correspondent families of modified sequential decision rules are developed. Within the developed families, the robustified sequential decision rules are constructed with the minimax risk criterion [10].

V. Conclusion

The approach is applied to analysis of COVID-19 incidence dynamics process in the Republic of Belarus to identify types of trajectories: growth, horizontal fluctuation, decrease [11]. Also cases of composite hypotheses can be treated with the discussed approach [9].

References

[1] N. Mukhopadhyay, and B. de Silva. Sequential Methods and Their Applications. New York, Marcel Dekker, 2009, 504 p.

[2] A. Wald. Sequential Analysis. New York, John Wiley and Sons, 1947, 212 p.

[3] T. Lai. "Sequential analysis: Some classical problems and new challenges," StatisticaSinica, 2001, vol. 11, pp 303–408.

[4] A. Y. Kharin. Robustness of Bayesian and Sequential Statistical Decision Rules. Minsk, BSU, 2013, 207 p. (In Russ.).

[5] P. Huber, and E. Ronchetti. Robust Statistics. New York, Wiley, 2009, 380 p.

[6] A.Y. Kharin, and D.V. Kishylau. "Robust sequential test for hypotheses about discrete distributions in the presence of "outliers"," Journal of Mathematical Sciences, 2015, vol. 205(1), pp 68–73.

[7] A. Kharin, T.T. Tu. "Performance and robustness analysis of sequential hypotheses testing for time series with trend," Austrian Journal of Statistics, 2017, vol. 46(3-4), pp 23–36.

[8] A.Y. Kharin, T.T. Tu. "On error probabilities calculation for the truncated sequential probability ratio test," Journal of the Belarusian State University. Mathematics and Informatics, 2018, no. 1, pp 68–76.

[9] A.Y. Kharin. An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses. Journal of Mathematical Sciences, 2017, vol. 227(2), pp 196–203.

[10] T.T. Tu, and A.Y. Kharin. Sequential probability ratio test for many simple hypotheses on parameters of time series with trend. Journal of the Belarusian State University. Mathematics and Informatics, 2019, no. 1, p 35–45.

[11] Y. S. Kharin, V. I. Malugin, V. A. Voloshko, O. V. Dzernakova, and A.Y. Kharin. Statistical forecasting of the dynaics of epidemiological indicators for COVID-19 incidence in the Republic of Belarus. Journal of the Belarusian State University. Mathematics and Informatics, 2020, no. 3, pp 36–50.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.