Sequential Monitoring Procedure for Change in Distribution

Marie Hušková, Ondřej Chochola

huskova@karlin.mff.cuni.cz, chochola@karlin.mff.cuni.cz

Katedra pravděpodobnosti a matematické statistiky,
Matematicko-fyzikální fakulta Univerzity Karlovy v Praze

A simple sequential procedure is proposed for detection of a change in distribution for dependent observations when a training sample with no change is available. Its properties under both null and alternative hypothesis are studied. Theoretical results are accompanied by a simulation study.

Introduction

We assume that the observations $X_1, \dots, X_n, \dots$ are arriving sequentially, X_i has the continuous d.f. F_i and the first m observations have the same distribution function, i.e., $F_1 = \dots = F_m = F_0$, where F_0 is unknown. $X_1, \dots, X_m$ are usually called training data. We are interested in testing the null hypothesis

$$H_0: \; F_i = F_0, \quad \forall i \geq m,$$

against the alternative

$$H_A: \exists k^* \text{ such that } F_i = F_0, \; 1 \leq i \leq m + k^*,$$
$$F_i = F^0, \; m + k^* < i < \infty, \quad F_0 \neq F^0.$$

Usually such testing problems concern only a change of finite dimensional parameter (see [2]). Preceding hypotheses were considered by e.g. [3] or [4], they use however rather strict assumptions. On the other hand we assume only the continuity of the distribution functions F_i in case of independent observations, in case of dependent ones certain type of dependency is assumed.

Our test procedure is described by the stopping rule:

$$\tau_{m,N} = \inf\{1 \leq k \leq N : \; |Q(m,k)| \geq c \, q_\gamma(k/m)\}$$

with standard understanding $\inf \emptyset := +\infty$ and either $N = \infty$ or $N = N(m)$ and $\lim_{m\to\infty} N(m)/m = \infty$.

The detector is choosen as

$$Q(m,k) = \frac{1}{\widehat{\sigma}_m \sqrt{m}} \sum_{i=m+1}^{m+k} (\widehat{F}_m(X_i) - 1/2), \quad k \geq 1,$$

where $\widehat{F}_m$ is an empirical distribution function based on $X_1, \dots, X_m$ and $\widehat{\sigma}_m$ is a suitable standardization based on $X_1, \dots, X_m$.

We use the boundary function

$$q_\gamma(t) = (1+t)(t/(1+t))^\gamma, \; t \in (0, \infty), \quad \gamma \in [0, 1/2),$$

with γ being a tuning parameter.

A positive constant c is chosen such that for fixed $\alpha \in (0,1)$ under H_0

$$\lim_{m\to\infty} P\big(\tau_{m,N} < \infty\big) \; = \; \alpha, \qquad (1)$$

We also require that under H_A

$$\lim_{m\to\infty} P\big(\tau_{m,N} < \infty\big) \; = \; 1. \qquad (2)$$

These requests mean that the test has asymptotically level α and asymptotical power one.

Main results

Here we formulate assertions on limit distribution of our test procedure under both null hypothesis as well under some alternative.

Under the null hypothesis we consider two sets of assumptions:

(H_1) $\{X_i\}_i$ are i.i.d. random variables with continuous distribution function F_0, $i = 1, 2, \dots$.

(H_2) $X_i = \mu + \kappa e_i$, $i = 1, 2, \dots$, where $\mu \in R^1$, $\kappa > 0$ and $\{e_i\}_i$ form a linear process: $e_i = \sum_{j=0}^{\infty} a_j \varepsilon_{i-j}$, where $\{a_j\}_j$ is a sequence of real numbers such that $|a_j| \leq C d^j$, $j = 0, 1, \dots$, for some positive C and $d \in (0,1)$ and $\{\varepsilon_i\}_{i=-\infty}^{\infty}$ are i.i.d. random variables with zero mean, unit variance and $E|\varepsilon_i|^4 < \infty$.

Next is the assertion on limit behavior of the functional of $Q(m,k)$ under the null hypothesis.

Theorem

(I) Let the sequence $\{X_i\}_i$ fulfill the assumption (H_1) and put $\widehat{\sigma}^2_m = 1/12$. Then

$$\lim_{m\to\infty} P\left(\sup_{1\leq k<\infty} \frac{|Q(m,k)|}{q_\gamma(k/m)} \leq x \right) = P\left(\sup_{0\leq t\leq 1} \frac{|W(t)|}{t^\gamma} \leq x \right) \qquad (3)$$

for all x, where $\{W(t);\, 0 \leq t \leq 1\}$ is a Wiener process.

(II) Let the sequence $\{X_i\}_i$ fulfill the assumption (H_2) and let, as $m \to \infty$, $\widehat{\sigma}^2_m - \sigma^2 = o_P(1)$, where

$$\sigma^2 = \frac{1}{12} + 2 \sum_{j=1}^{\infty} cov\{F_0(X_1), F_0(X_{j+1})\}. \qquad (4)$$

Then (3) holds true.

Theorem provides approximation for the critical value c so that the test procedure fulfills (1) under the null hypothesis (H_1 or H_2), i.e., c is the solution of the equation

$$P\left(\sup_{0\leq t\leq 1} \frac{|W(t)|}{t^\gamma} \leq c \right) = 1 - \alpha. \qquad (5)$$

For the alternative hypothesis we require that $\int F_0(x) dF^0(x) \neq 1/2$. Then it can be shown that

$$\sup_{1\leq k<\infty} \frac{|Q(k,m)|}{q_\gamma(k/m)} \xrightarrow{P} \infty, \quad \text{as } m \to \infty.$$

This ensures that the requirement (2) is met.

Proofs of both assertions can be found in [1]. Here we just sketch a basic idea of the proof of Theorem. We need to show that the limit distribution of $\{V_m(t),\, t > 0\}$, where $V_m(t) = \frac{1}{\sqrt{m}} \sum_{i=m+1}^{m+\lfloor mt \rfloor} (\widehat{F}_m(X_i) - 1/2)$ is the same as of $\{Z_m(t),\, t > 0\}$, where $Z_m(t) = \frac{1}{\sqrt{m}} \Big(\sum_{i=m+1}^{m+\lfloor mt \rfloor} (F_0(X_i) - 1/2) - \frac{k}{m} \sum_{j=1}^{m} (F_0(X_j) - 1/2) \Big)$. Moreover a process $\{\frac{1}{\sqrt{m}} \sum_{i=m+1}^{m+\lfloor mt \rfloor} (F_0(X_i) - 1/2),\, t > 0\}$ converges to a Gaussian process in a certain sense as $m \to \infty$ and $\frac{1}{\sqrt{m}} \sum_{j=1}^{m} (F_0(X_j) - 1/2)$ converges in distribution to $N(0, \sigma^2)$, where σ^2 is the same as in (4).

As Theorem indicates, in case of dependent observations we need a consistent estimator of σ^2. We use

$$\widehat{\sigma}^2_m = \widehat{R}(0) + 2 \sum_{k=1}^{\Lambda_m} w(k/\Lambda_m) \widehat{R}_m(k), \;\text{ where} \qquad (6)$$

$$\widehat{R}_m(k) = \frac{1}{n} \sum_{i=1}^{n-k} (\widehat{F}_m(X_i) - 1/2)(\widehat{F}_m(X_{i+k}) - 1/2) \text{ and}$$

$$w(t) = 1I\{0 \leq t \leq 1/2\} + 2(1-t)\{1/2 < t \leq 1\}$$

is a weight function. Under suitable assumptions on Λ_m this estimator is consistent.

Simulations

Here we report selected results of a simulation study that was performed in order to check the finite sample performance of the monitoring procedure. All results are for the level $\alpha = 5\%$. The asymptotic critical values, i.e. the values obtained from (5), were used.

The parameters chosen:

- Length of training data: $m = 50, 100, 500$
- Observations: independent or AR(1) process with $\rho = 0.2, 0.4$
- Distribution of innovations: t_4 and demeaned LN(0,1)
- Change point: $k^* = 0$ i.e. the start of the monitoring
- Tuning constant: $\gamma = 0, 0.25, 049$

Since the procedure make use of an empirical distribution function it is convenient also for distributions with heavier tails. Therefore we do not report results for e.g. normal distribution, even though they are better than the reported ones. The estimate $\widehat{\sigma}^2_m$ is set to 1/12 for independent observations and is calculated according to (6) for dependent ones.

The empirical sizes of the procedure under the null hypotheses are reported in the table. We can see that for independent observations the level is kept and the prolongation of the training period has no significant effect. This is not the case when for estimating σ^2 the formula (6) is used, since there we need more data to estimate it precisely enough and therefore the prolongation will bring the empirical size closer to the required level. Similar holds for dependent observations. Since we will later examine an early change, we are mostly interested in γ close to 1/2. For $\gamma = 0.49$, the results are satisfactory even for the dependent observations.

		t$_4$			LN		
ρ	$m \setminus \gamma$	0	0.25	0.49	0	0.25	0.49
	50	4.4	4.3	1.7	4.3	4.1	1.7
0	100	4.7	4.3	2.0	4.7	4.3	2.0
	500	4.2	4.4	2.8	4.4	4.5	3.0
	50	8.6	8.7	4.6	9.0	8.8	4.6
0.2	100	6.6	6.4	3.8	7.5	7.5	3.9
	500	5.0	5.3	3.5	5.2	5.6	3.8
	50	10.3	10.4	5.2	11.0	10.9	5.4
0.4	100	9.0	9.3	4.8	8.9	8.8	4.2
	500	6.7	7.2	4.9	7.2	7.4	5.0

Now we focus on alternatives. Since $k^* = 0$, the stopping time equals the detection delay. The table presents a summary of stopping times for a unit change in mean.

		$\rho = 0$			$\rho = 0.2$			$\rho = 0.4$		
	$\setminus m$	50	100	500	50	100	500	50	100	500
	1st Qu.	8	9	10	37	34	32	56	47	42
	Median	11	13	13	63	50	42	141	78	60
t$_4$	Mean	12	15	14	120	61	45	234	125	66
	3rd Qu.	15	18	17	123	71	56	500	140	83
	Max.	52	54	46	500	500	168	500	500	365
	1st Qu.	9	10	10	19	19	19	38	33	32
	Median	14	13	13	34	27	24	113	58	45
LN	Mean	23	15	13	90	38	26	230	116	51
	3rd Qu.	23	18	16	76	41	31	500	122	62
	Max.	500	75	30	500	500	67	500	500	324

For independent observations the prolongation of the training period leads mainly to reducing extremes of the delay (more clearly visible for smaller amount of change), whereas for dependent ones the impact of increased m is overall significant. With a growing dependence amongst the data, the performance of the procedure is worsening. However the results for $m = 500$ are satisfactory even with an autoregressive coefficient $\rho = 0.4$. For shorter training period, the change was not detected in some replications (stopping time equals 500 then).

We performed also tests for change in variance and type of distribution. The detection delays there are longer, however the procedure is still able to detect the change. For more details see [1].

Acknowledgement

The work has been supported by grants GAČR 201/09/J006, MSM 0021620839 and LC06024.

References

[1] Hušková M., Chochola O. Simple sequential monitoring procedure for change in distribution, *to be published*

[2] Horváth L., Hušková M., Kokoszka P., and Steinebach J. Monitoring changes in linear models. *J. Stat. Plann. Inference*, 126:225–251, 2004.

[3]Lee S., Lee Y. and Na O. Monitoring distributional changes in autoregressive models. *Commun.Statist.Theor.Meth.* 38:2969–2982, 2009.

[4] Mukherjee, A. Some Rank-Based Two-Phase Procedures in Sequential Monitoring of Exchange Rate. *Sequential Analysis*, 28: 137–162, 2009.