International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Comparison Online Monitoring Method of Correlated High-Dimensional Data Streams

Shuixian Bian

Tianjin University of Technology and Education No.1310,Dagu South Road, Hexi District, Tianjin 300222, China

Abstract: ***In this article we consider four charts(max and sum of cumulative sum,higher criticism statistic and goodness-of-fit test statistic)for monitoring related high dimensional data streams to find the alarming time as quickly as possible after the mean shift and find the ideal chart in different condition. We use the robust of different one-side statistics*** $\mathrm{D_{max}}$, $\mathrm{D_{sum}}$, $\mathrm{D_{HC}}$, $\mathrm{D_{GOF}}$ ***to find the best chart in sparse case and dense case.From the results analysis,we can gain that the goodness-of-fit test has the best efficient from balancing the power and robust in theory.***

Key words: CUSUM, Higher Criticism Statistic, Goodness-Of-Fit Test Statistic, One-side Statistic, Order statistic.

1. Introduction

In several years, high dimensional data streams become more and more popular in industrial application, thereby how to select the proper chart and monitor them to reduce the fraction defective and waste of the products,which has become very important. In fact, the data streams are related. Before now, some has an assumption that date streams are independent, comparing to the max ,sum, HC and GOF monitored random variable to choose the proper chart based on likelihood ratio test and test the GOF is best to other three charts. Now we test it still hold by one-side statistic.

2. Model

For the correlation of the date streams,model is assumed by

$$\mathrm{X}_t = U + \mathrm{Z}_t \tag{1}$$

Where the mean vector U is nonrandom and sparse, Z_t and X_t are $p\times 1$ -dimensional vector,in control, $\mathrm{Z}_t \overset{\text{iid}}{\sim} N(0,\Sigma)$

$$,\quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}_{p\times p}$$

is the covariance matrix of $\{Z_{1t}, Z_{2t}, \cdots, Z_{pt}\}$.so we can standardize it to that $Y_t \overset{\text{iid}}{\sim} N(0, I_p)$, after that we only consider Y_t.

3. The brief descriptions of the methods

Sum and Max of the CUSUM statistic. At time point t,we can get p correlative observations $Y_t = (Y_{1t},\ldots,Y_{pt})$,each sample of which is from $N_p(0,I_p)$ if that is not affected otherwise which is from $N_p(u_n, I_p)$ and whose probability affected is ε_n. Based on the CUSUM statistic

$$S_k(t) = \max\{0, S_k(t-1) + u_k(Y_{kt} - u_k/2)\} \tag{2}$$

Where u_k is the constant mean value of each stream we given.

We can obtain that

$$T_{\max} = \inf\{t : \max_{k=1,\ldots,p} S_k(t) \ge L\} \tag{3}$$

$$T_{sum} = \inf\{t : \sum_{k=1}^{p} S_k(t) \ge L\} \tag{4}$$

Where L is the 95%-upper-fractile of monitoring from standard normal distribution.

Their one-side statistic are $\mathrm{D_{max}}$, $\mathrm{D_{sum}}$,which are used to illustrate the robust.

Higher Criticism Statistic. From[9]higher criticism statistic HC_n^* is defined as following

$$\mathrm{HC}_n^* \equiv \max_{1\le k\le p} \mathrm{HC}_{n,k},\quad \mathrm{HC}_{n,k} = \frac{\sqrt{p}(k/p - p_{(k)})}{\sqrt{p_{(k)}(1-p_{(k)})}} \tag{5}$$

Where

$p_k = 1 - \Phi(Y_k) \equiv \bar{\Phi}(Y_k)$, $\Phi(Y_k) = P\{N(0,1) > Y_k\}$ is the cdf of the standard normal distribution and $p_{(1)} < p_{(2)} < \cdots < p_{(p)}$ are the order statistics of p values.

Therefore the stopping time is obtained that

$$T_{HC} = \inf\{t : \mathrm{HC_n^*}(t) \ge \mathrm{L}\} \tag{6}$$

Whose one-side statistic is given as $\mathrm{D_{HC}} \equiv \max_{1\le k\le p} \mathrm{HC}_{n,k}$.

Goodness-of-fit Test Statistic. Depend on the higher criticism, we can suggest the one-side statistic of GOF like following expression

$$D_{GOF} = \sum_{k=1}^{p}\left\{\log\left[\frac{[\Phi(Y_{(k)})]^{-1}-1}{(p-1/2)/(k-3/4)-1}\right]\right\}^2 \times I_{\{\Phi(Y_{(k)})>(k-3/4)/p\}} \tag{7}$$

Where $I(\cdot)$ is the indicator function, $Y_{(1)} < Y_{(2)} < \cdots < Y_{(p)}$ is the order statistic. So replacing Y_i by $S_k(t)$, $k = 1,\ldots,p$ in (7) could have equation

International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

$$W_t = \sum_{i=1}^{p}\left\{\log\left[\frac{U_{(i)}^{-1}(t)-1}{(p-1/2)/(i-3/4)-1}\right]\right\}^2 \times I_{\{U_{(i)}(t)>(i-3/4)/p\}} \qquad (8)$$

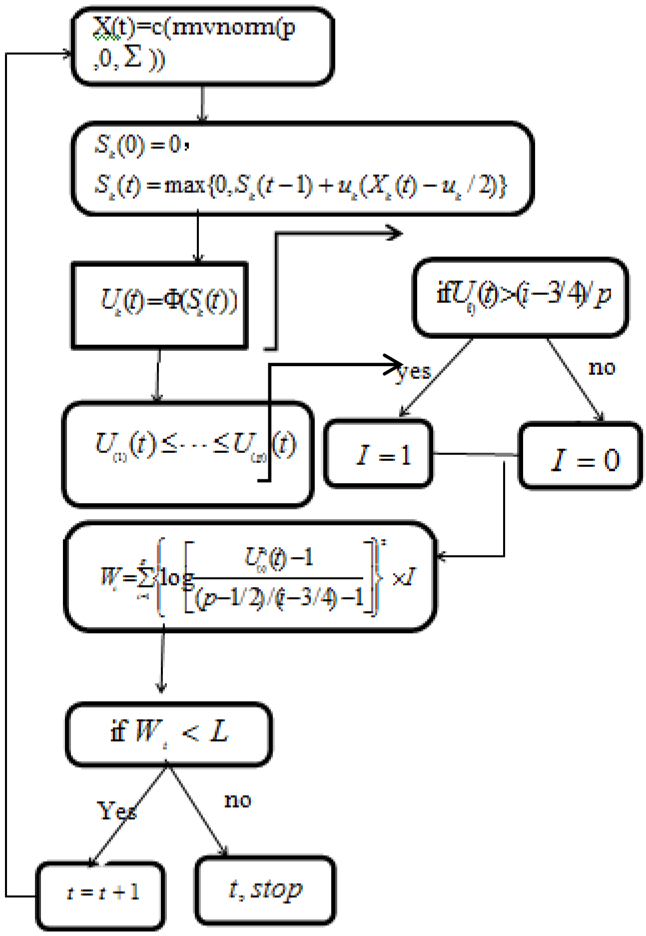
Where

$$U_{(1)}(t) \le \cdots \le U_{(p)}(t)$$

is the order statistic of $\{U_{(1)}(t),\ldots,U_{(p)}(t)\}$, and $U_{(i)}(t) = H_t(S_i(t);\mu_i)$, $H_t(\cdot;\mu)$ denote the cdf of $S_i(t)$ about supposed parameter μ in control station.then the stopping time is defined as following $T_{new} = \inf\{t : W_t \ge L\}$.

4. Performance Comparison

To understand clearly, we can give a flow chart by a sample of application of GOF statistic in control state.



Where X(t) denotes monitoring p values at time point t , u_k is the given mean value of kth date stream, $I(.)$ is the indicator function. $\Phi(S_k(t))$ is the ecdf of $S_k(t)$, L is a control limit chosen to achieve a specific value of IC average run length (ARL)and positive.

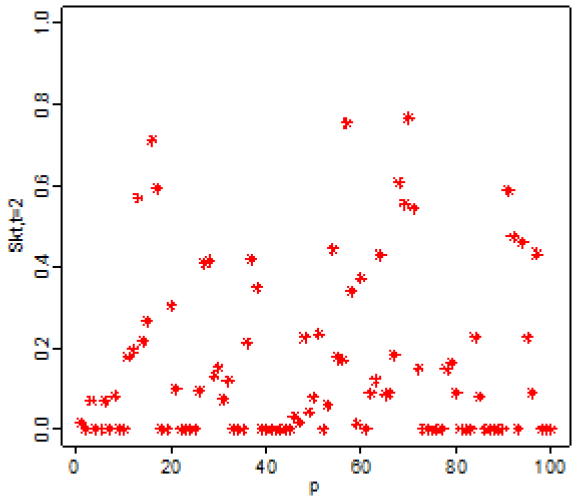


Figure1, when t is from 0 to 200, ordinate denote $S_k(t)$ change with $u_k = 0.2$ given in control state.

From that, we could extend it to p-dimensions.

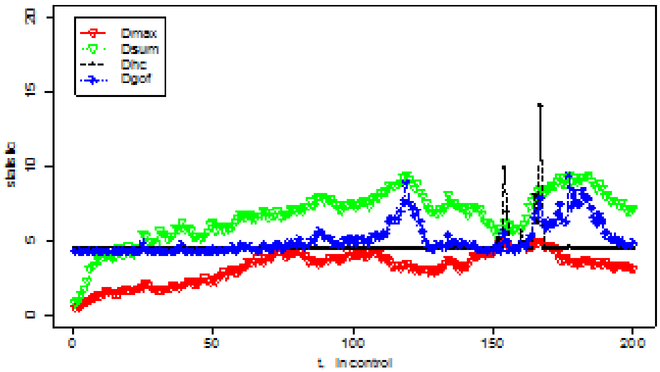


Figure2,when $p = 100$,one-side statistics of four charts in control.

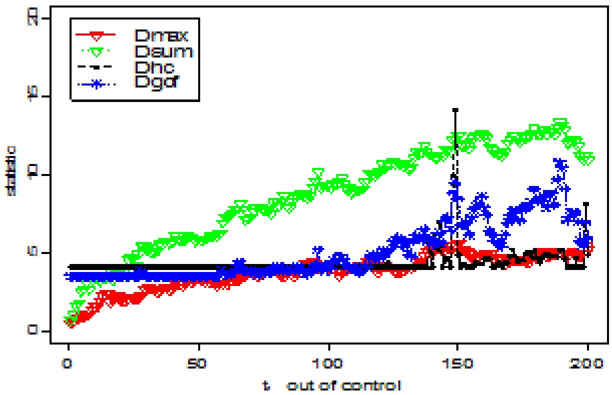


Figure3,when $p = 100$,one-side statistics of four charts out of control(u=0.05).

Step 1:we assume the ARL_0 in control is given, the first type error $\alpha = \frac{1}{ARL_0}$ could be sure. we get p dimension correlative random values $X_{p\times n} = (X_1,\ldots,X_p)^T$ (n is the times of simulating based Monte Carlo simulation. Since the covariance matrix is always symmetric and positive definite, the translation is practicable.

Step 2: Simulating n times to calculate D_{max} , D_{sum} , D_{HC} , D_{GOF} in different time points in control, rank them and find $L_{\max}, L_{sum}, L_{HC}, L_{GOF}$ in the quantile $(1-\alpha)n$.

Step 3: Once the statistic exceed the L, the corresponding time is the stopping time T, meanwhile we can get the confirm the FDR β , then $ARL_1 = \dfrac{1}{1-\beta}$.

Step 4: Then online monitoring ,we compute $D_{\max}$, D_{sum} , D_{HC} , D_{GOF} at each time t out control ,respectively. Meanwhile we can compute the rate of efficient alarm for presumptive change point time τ to measure the methods' robust in different correlation matrices.

Table 1,p=100,warning times T of four statistics

μ $T(sd)$	T_{sum}	$T_{\max}$	T_{hc}	T_{new}
0.00001	65.5822 (40.5556)	70.22418 (39.5857)	67.70964 (41.5166)	69.044 (36.6028)
0.0001	65.06337 (40.6159)	70.0807 (39.1541)	64.95132 (42.0102)	67.374 (37.0121)
0.001	64.02635 (40.4058)	68.23397 (39.9023)	66.74512 (41.4918)	67.472 (36.7714)
0.01	54.26584 (37.4622)	65.1187 (39.1914)	58.49866 (39.4505)	62.253 (36.7520)
0.02	44.33917 (32.1257)	61.30632 (38.6764)	52.36828 (36.5411)	54.591 (32.1812)
0.05	25.36329 (19.0507)	51.1201 (33.0558)	37.31158 (27.3399)	36.205 (28.6594)
0.1	13.5869 （9.60038）	37.81227 （25.0808）	21.52461 （14.9779）	22.94553 （13.7824）
0.2	6.95375 （4.48577）	25.0631 （14.7225）	10.37244 （6.57241）	11.67791 (6.267173)
0.5	2.832512 (1.597044)	12.6426 (6.967676)	3.936743 (2.188187)	4.679515 (2.27336)

5. Conclusion

From Table 1,we can obtain that: with the drift gradually increase, the sensitivity of T_{sum} is better, but T_{new} is the best for the small drift, meanwhile the robustness is the best from others.

Good-Of-Fit test statistic may be not always best in all case, but never be worst. In general, the method is better.Of course, we could choose the best method in different case, so that the control chart is the most efficient and reduce the error ratio.

References

[1] Moustakide,G.V., Optimal Stopping Times for Detecting Changes in Distributions[J]. The Annals of Statistics. (1986).
[2] Mei, Y., Efficient Scalable Schemes for Monitoring a Large Number of Data Streams[J]. Biometrika, 2010(97), 419-433.
[3] Hall, P., and Jin, J. Innovated Higher Criticism for Detecting Sparse Signals in Correlated Noise[J]. The Annals of Statistics, 2010(38), 1686-1732.
[4] Zhang, J. Powerful Goodness-of-Fit Tests Based on Likelihood Ratio[J]. Journal of the Royal Statistical Society, Series B, 2002(64): 281-294.
[5] Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. Detecting Simultaneous Change-Points in Multiple Sequences[J]. Biometrika, 2010(97): 631－645.
[6] Zou, C., Jiang, W., and Tsung, F. A LASSO-Based SPC Diagnostic Framework for Multivariate Statistical Process Control[J].Technometrics, 2011(53):297-309.
[7] Zou, C., and Qiu, P. Multivariate Statistical Process Control Using LASSO[J]. Journal of the American Statistical Association, 2009(104): 1586-1596.
[8] Brook, D., and Evans, D. A. An Approach to the Probability Distribution of CUSUM Run Length[J]. Biometrika, 1972(59),:539－549.
[9] Donoho, D., and Jin, J. Higher Criticism for Detecting Sparse Heterogeneous Mixtures[J]. The Annals of Statistics, 2004(32): 962-994.
[10] Qiu, P., and Xiang, D. Univariate Dynamic Screening System: An Approach for Identifying Individuals With Irregular Longitudinal Behavior[J]. Technometrics, 2014 (56): 248-260.

Author Profile

Shuixian Bian is reading the M.S. degrees in College of Science from Tianjin University of Technology and Education .