A two-stage monitoring scheme for high-dimensional Poisson data

Huantong Na[*1,a], Xuemin Zi[1,b]

[*a]1812473526@qq.com, [b]zi_xuemin@aliyun.com

[1]Tianjin University of Technology and Education, Tianjin, China

Abstract: In the field of industrial quality control, when monitoring continuous data is mainly considered, it is also necessary to monitor discrete data, among which high-dimensional Poisson distribution data is very common in modern manufacturing. At present, a large number of the literature proposed to use of statistical process technology to monitor high-dimensional data, but the scheme for monitoring high-dimensional Poisson data in the existing literature is rare, and the few proposed methods for high-dimensional Poisson data to monitor the false discovery rate (FDR) of data were given, which is popular statistic for constructing the control chart. In order to effectively monitor the high-dimensional Poisson data, this paper extended the methodology of (Li, 2018) proposes a two-stage monitoring algorithm, which uses the CUSUM control chart to control the in-control (IC) average run length (ARL) in the first stage, and uses pointwise FDR to control the Type-I error rate in the second stage. In this paper, numerical simulation is used to realize the monitoring process and the performance demonstrates the efficiency and robustness of the proposed procedure.

Keywords: high-dimensional Poisson data, CUSUM control chart, ARL, FDR

1 INTRODUCTION

In recent years, due to the continuous emergence of "big data", the formation of data collection and computing technology has been promoted, and high-dimensional data monitoring has also become a highly desired topic in the field of statistical process control (SPC). The complexity of product technology and quality monitoring is increasing with the emergency of revolutionary procedures, which are totally different from the conventional one, and discrete enterprises such as machinery, automobiles, home appliances, clothing, and so on need to process various parts provided by different suppliers, and the process will produce a large number of high-dimensional discrete data. For example, in the aerospace manufacturing industry, it is necessary to monitor the number of defects in a space part, and if the process cannot be effectively monitored and defects are found in time, it will cause huge losses and reputational impact on the enterprise. Moreover, at present, discrete enterprises have a huge workload of data collection, inspection and maintenance, and labor requires a lot of time to make judgments, which increases production costs. At this point, statistical process control chart is considered as an effective tool to monitor high-dimensional discrete data in the realistic scenarios.

When people monitor this discrete data, they assume that the numbers of product defects or nonconformities follow the Poisson distribution, which is one of the most important discrete distributions to describe the quantity of defects of product. The most common method of

EMIS 2023, February 24-26, Chengdu, People's Republic of China

DOI 10.4108/eai.24-2-2023.2330695

monitoring Poisson distribution data is to use the C-chart and U-chart from the Shewhart chart. Use the C-chart when the sample size is the same; when the sample size is different, the U-chart is used. But the downside is that when the quantity changes more than expected, there is an over-discretization phenomenon, making it uncontrolled. In addition that the Shewhart chart has the defect of insensitive detection of small deviations in the process. To compensate for this shortcoming, Page designed the CUSUM chart in 1954. The CUSUM chart is a quality monitoring process that calculates the cumulative sum of deviations between each sample value and the target value, which solves for small shifts in the process.

In the industrial production process, it is generally necessary to monitor multiple attributes of a product at the same time, so multiple control charts are required. For CUSUM charts, the control limit is determined by fixing the average run length (ARL) of the population in the in-control (IC) state. IC ARL refers to the average number of samples taken from the start of the detection to the time it stops when the production process is in the IC state, that is, the frequency at which the monitor wants false alarms to occur when the process is in the IC state. But when the number of charts is large, the capabilities of such a scheme are quite low.

In this paper, we mainly focus on the discrete data with Poisson distribution. In order to monitor Poisson distribution data, [9](Suvimol and Chananet, 2022) proposed a display formula based on the generalized Poisson distribution (MAGP) moving average chart and performed performance testing of the MAGP chart. The results show that the MAGP chart is superior to the Shewhart control chart (CGP) chart based on the generalized Poisson distribution in detecting small shifts. [11](Xiao and Zi, 2021) propose to monitor multivariate Poisson distribution data using a multivariate exponentially weighted moving average (MEWMA) control chart. The results show that the MEWMA chart can be well controlled for the multivariate Poisson data with different parameter shifts. [10](Alevizakos et al., 2021) propose a three-exponentially weighted moving average control chart (PTEWMA chart) for monitoring the Poisson process, which is compared with the existing PEWMA and PDEWMA charts. The results show that the detection ability of the PTEWMA plot becomes stronger and stronger as the value increases.

False discovery rate (FDR) is widely used in control charts. [1](Benjamini and Hochberg, 1995) proposed a FDR control method to control the local false alarm rate. FDR is the expected value of the proportion of all hypothesis null but rejected as a proportion of all rejected hypotheses in all hypothesis tests. It is also the number of false alarms that monitors can tolerate when OC data is identified. This paper theoretically demonstrated that controlling FDR is equivalent to controlling the Type-I error rate when all null hypotheses are true. At present, FDR has been widely used in most fields. For example, for change point detection of faults or fault elimination in multi-stream data monitoring, see [7](Sehgal et al., 2010) and for literature considering multiple hypothesis testing problems and microarray studies, see [5](Li and Tung, 2009). The inadequacy is that this monitoring methods does not make use of global information. Even if the local false alarm rate can be well controlled on each data stream, but when the number of data streams is large, the global false alarm rate may be serious, resulting in large detection delays and increasing production.

In order to effectively monitor the high-dimensional Poisson data streams, we propose a two-stage monitoring algorithm based on [4](Li, 2018) combining CUSUM control chart with FDR. The first stage determines whether there is an anomalous data stream by monitoring the global FDR. If the first stage determines that there is at least one out-of-control (OC) data stream, the local FDR of the OC data streams is monitored and the locations of the OC data streams are determined in the second stage. In the product production process, the two-stage monitoring

algorithm proposed in this paper can not only satisfy the frequency of false alarms expected by the monitor when the process is IC, but also satisfy the number of false alarms that the monitor can tolerate when the OC data is identified. The article is organized as follows. Section II, a two-stage monitoring algorithm for high-dimensional Poisson data is proposed. Section III gives a numerical simulation to illustrate. Section IV gives a summary and concluding remarks.

2 MATERIALS AND METHODS

When monitoring high-dimensional data streams, there are generally two single-stage monitoring methods. The first is to identify the location of the OC data stream through the control point FDR. The application of the point FDR to different locations indicates that there are different problems in the system. Therefore, different discriminant rules are needed, see [5](Li and Tsung, 2009). The second is to identify the location of the OC data streams by controlling the global FDR, which is suitable that where different OC data streams indicate that only one problem will occur in the entire system. Therefore, the discrimination rules are the same and there is no need to identify the existence of OC data stream, see [12](Zou et al., 2015). These two methods have not been combined in the past research literature, and cannot satisfy the flexibility of selecting a single-point FDR at user level. In order to solve these limitations, [4](Li, 2018) proposed a two-stage monitoring procedure, which links two single-stage monitoring methods. This method is based on high-dimensional normal data streams. Select the global monitoring statistic proposed by [12](Zou et al., 2015) as the first stage statistic. According to the *p*-value formula based on normal distribution CUSUM statistic proposed by [3](Grigg and Spiegelhalter, 2008), the hypothesis test and monitoring of global FDR are carried out. If there is at least one OC data stream in the first stage, move to the second stage. In the second stage, the standardized CUSUM statistic of the OC data stream is the statistic of the second stage, and the location of the OC data stream is identified by the control point FDR. This article extends this idea to high-dimensional Poisson data streams. The normal distribution is continuous but the Poisson distribution is discrete, so it cannot be directly adopted. Therefore, it is necessary to transform the Poisson data to approximately satisfy the normal properties approaches. In this paper, a proximity method is proposed to calculate the *p*-value of the Poisson distribution CUSUM statistic.

To begin with, denote the numbers of IC data streams at time *t* by $m_{0,t}$ and $m_{1,t} = m - m_{0,t}$ represent the number of OC data streams, both of which are unknown. Let R_t be the number of rejections of the null hypothesis, $m - R_t$ be the number of null hypotheses accepted, V_t be the number of false findings of Type-I errors, $m_{0,t} - V_t$ be the number of accepted null hypotheses true, S_t be the number of correctly rejected hypotheses, and $m_{1,t} - S_t$ be the number of Type-II errors, which are unknown. Then the pointwise FDR at time *t* is symbolized by if $R_t > 0$, then

$$FDR_t = E(Q_t) = E\left(\frac{V_t}{R_t}\right); \text{ If } R_t = 0 \text{ then}$$

$FDR_t = 0$.

The overall FDR is the sum of the point FDR. It can be written as

$GFDR_t = E\left(\frac{\sum_{t=1}^{T} V_t}{\sum_{t=1}^{T} R_t \vee 1}\right)$. If the point FDR is controlled at the level of α / T, then the global FDR is controlled at the level of α for every *t*.

To monitor the global FDR and the point FDR, the monitoring problem is transformed into a two-stage hypothesis testing problem. There are *m* dimensional data streams in the process, where the *i*th data stream observed at time *t* by $X_{i,t}$, $i = 1, 2, \ldots m$ and $t = 1, 2, \ldots$. Without loss of generality, it is assumed that $X_{i,t}$ are independent of each other. When the process is IC, the distribution of $\{X_{1,t}, X_{2,t}. \ldots\}$ follows the Poisson distribution, denoted by $F_{i,0}$, $i = 1, \ldots, m$. When the process becomes OC from a certain point $t_i, t_i \in [1, t]$, the distribution of $\{X_{i,1}, X_{i,2}, \ldots, X_{i,t_{i-1}}\}$ follows $F_{i,0}$, the distribution of $\{X_{i,t_i}, \ldots, X_{i,t}\}$ follows $F_{i,1}$. Then multiple hypothesis tests can be established for $i = 1, \ldots, m$, as follows:

$H_{0,i,t} : X_{i,1}, X_{i,2}, \ldots, X_{i,t} \sim F_{i,0}$,

versus

$H_{1,i,t} : \exists t_i \in [1, t]$ such that

$X_{i,1}, X_{i,2}, \ldots, X_{i,t_{i-1}} \sim F_{i,0}$ and

$$X_{i,t_i}, \ldots, X_{i,t} \sim F_{i,1} \tag{1}$$

where $F_{i,0}$ is the IC distribution, $F_{i,1}$ is the OC distribution, and t_i is the change point of the *i*th data stream.

If $H_{0,i,t}$ is accepted, it indicates that all data streams are IC. If $H_{0,i,t}$ is rejected, it indicates that some data streams are OC at time *t*. In this case, move to the second stage to identify where data streams are OC.

The existing method is to test the data streams over time, but this single method cannot quickly determine whether there is OC data stream and where OC data stream is located. In order to break this limitation, this paper proposes a two-stage monitoring algorithm based on [4](Li, 2018). Since this paper is aimed at high-dimensional discrete Poisson data streams, and the *p*-value calculation method proposed by [3](Grigg and Spiegelhalter, 2008) is applicable to continuous normally distributed CUSUM data, it cannot be directly adopted. In the second II, a proximity

method is proposed to calculate the *p*-value of high-dimensional Poisson CUSUM statistic. The components of the two-stage algorithm are as follows: global information is extracted from all data streams for global FDR monitoring and hypothesis testing in the first stage. If Eq. (1) is true, there is no OC data stream. If Eq. (1) is rejected, then the second stage is entered. In the second stage, construct local statistics and test data streams that occur when Eq. (1) is rejected to determine the location of OC data streams.

The two-stage monitoring algorithm proposed in this paper uses the control limit of each stage to make the algorithm satisfy the user's requirements for IC ARL and Type-I error rate. Therefore, the algorithm can satisfy the frequency of false positives expected by the user when the process is IC and the number of false positives that the user can tolerate when identifying OC data.

2.1 Monitoring The Global FDR

(Rossi et al., 1999)[6] proposed a standardized CUSUM chart based on the Poisson distribution and compared three methods for converting Poisson data to approximately normal data. The first transformation is

$$Z_{1,i} = \frac{X_i - n_i l_0}{\sqrt{n_i l_0}}, i = 1, 2, \mathbf{L}$$

This conversion is based on the asymptotic normality of the observed number X_i, with an approximate control mean $n_i l_0$ and an approximate standard error $\sqrt{n_i l_0}$. The second transformation is called the square root transformation,

$$Z_{2,i} = 2\left(\sqrt{X_i} - \sqrt{n_i l_0}\right), i = 1, 2, \mathbf{L}$$

It has an approximate control mean $\sqrt{n_i l_0}$ and an approximate standard error $1/2$.The second transformation uses a normalized transformation, which stabilizes the variance. The third transformation is the average transformation of the first two transformations, the half-sum transformation, and the expression is

$$Z_{3,i} = \frac{X_i - 3n_i l_0 + 2\sqrt{X_i n_i l_0}}{2\sqrt{n_i l_0}}, i = 1, 2, \mathbf{L}$$

The control mean of this transformation results is approximately equal to 0, and the standard error is approximately equal to 1.

(Rossi et al., 1999)[6] used a table by (Ewan and Kemp, 1960) summarizing the performance of CUSUM charts of Poisson variables as a criterion for comparing three transformations. They found that the ARL values produced by the half-sum transformation were closest to those calculated by [2](Ewan and Kemp, 1960).

Therefore, the third transformation is the most efficient transformation. In construction, a half-sum transformation is taken to approximate Poisson data into standard normal data.

In the first stage, Poisson distribution data is generated and approximated as standard normal

distribution data using the half-sum transformation method proposed by [6](Rossi et al., 1999), is defined as

$$Z_{3,i} = \frac{X_i - 3n_i\lambda_0 + 2\sqrt{X_i n_i \lambda_0}}{2\sqrt{n_i\lambda_0}}, i = 1, 2, \ldots. \tag{2}$$

The mean of the OC data at $\lambda = \lambda_1$ is approximated

$$E\left[Z_{3,i};\lambda_1\right] = \frac{n_i\lambda_1 - 3n_i\lambda_0 + 2n_i\sqrt{\lambda_0\lambda_1}}{2\sqrt{n_i\lambda_0}}, i = 1, 2, \ldots. \tag{3}$$

Under the assumption of Poisson distribution, the most suitable global statistic for each data stream is the CUSUM statistic, which is defined as

$$\begin{cases} C_{i,0}^{+} = 0 \\ C_{i,t}^{+} = \max\left(0, S_{1,i-1} + Z_{3,i} - \dfrac{E\left[Z_{3,i};\lambda_1\right]}{2}\right), i = 1, 2, \ldots. \end{cases} \tag{4}$$

(Zou et al., 2015)[12] proposed a global statistic that performs well in identifying OC data streams and does not require a prior knowledge of the number of OC data streams. Hence, the statistic proposed (Zou et al., 2015)[12] is defined as the global statistic of the first stage

$$G_t = \sum_{i=1}^{m} \left\{ \log\left[\frac{\left(1 - p_{(i),t}\right)^{-1} - 1}{(m - 1/2)/(i - 3/4) - 1} \right] \right\}^2 I_{\left\{p_{(i),t} < 1 - (i - 3/4)/m\right\}}, \tag{5}$$

where $I_{\{A\}}$ represents the indicator function, taking 1 if *A* is true and 0 otherwise.

For calculating *p*-value quickly, (Grigg and Spiegelhalter, 2008)[3] proposed a close-form formula to approximate the steady-state *p*-value which the CUSUM statistic is generated from normal data, but this formula does not apply to calculate *p*-value of the CUSUM statistic generated from Poisson data. So this paper propose a proximity value method. The proximity value method means that the CUSUM statistic is generated from the normal distribution each time and arranged from small to large, denoted by $c_{i,t}$. The corresponding *p*-value is calculated according to (Grigg and Spiegelhalter, 2008).[3] Denote the *p*-value of by $P_{i,t}$. $c_{i,t}$ and $P_{i,t}$ is used as the standard value. Then regenerate CUSUM statistic by $Z_{3,i}$, differing from to choose which is the closest, and take the corresponding *p*-value as the Poisson distribution CUSUM statistic, denoted by $p_{i,t}$.

The CUSUM statistic generated over time is hypothesis tested according to Eq. (1). When all data streams are IC, the user's expected false alarm rate is the control limit *h* of the first stage, where *h* can be calculate by Monte Carlo simulation to meet IC ARL. If $G_t > h$, there is at least

one OC data stream and moves to the second stage. Otherwise, all data streams are IC.

2.2 Monitoring The Local FDR

Represents $W_{i,t}$ the local statistic of the second stage, which is equal to G_t standardized value in the case of $G_t > h$ in Eq. (1).

First, calculate the control limit of the second stage c_h to satisfy the Type-I error rate, the steps are as follows:

Step 1: Given c_h , use G_t to calculate the proportion of $W_{i,t} > c_h$ in the case of $G_t > h$ and express it by $\hat{p}$;

Step 2: Repeat the first step enough times (e.g. 2000 times), record $\hat{p}$ each time and average value;

Step 3: Record the value obtained in Step 2 as c_h .

If $W_{i,t} > c_h$, so the *i*th data is out of control. Otherwise, there are no OC data streams.

The first stage of monitoring flow chart is shown in Figure 1.

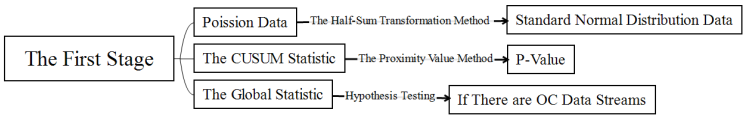

Figure 1: The first stage of monitoring flow chart

The second stage of monitoring flow chart is shown in Figure 2.

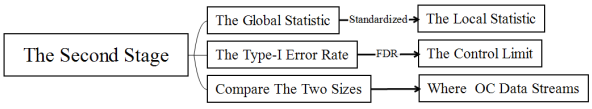

Figure 2: The second stage of monitoring flow chart

3 RESULTS & DISCUSSION

In the simulation, we assume that when the process finds that an OC data stream exists, it will not stop monitoring. The process stops monitoring only when all OC data streams are detected.

Simulations have developed different combinations of l_0 and l_1. This is shown in Table 1.

Table 1: Different combinations of l_0 and l_1.

l_0	l_1
6	7
6	8
6	9
10	12
10	14
10	18
10	20

In the two-stage monitoring numerical simulation, we take $l_0 = 10, l_1 = 12$ as an example, and other combinations have obtained similar conclusions. We choose the number of normal data to 200, the number of Poisson distributions to 100. The required IC ARL is set to 200, 500, 1000, and 10000. In addition to the above settings, PCER is divided into three cases: PCER=0.01, 0.03, and 0.05. Then the control limit h of the first stage and the control limit c_h of the second stage are given in Table 2.

Table 2： The control limits used in the two-stage procedure when m=100.

IC ARL	h	FDR	c_h	IC ARL	h	FDR	c_h
200	12.20215	0.01	0.9999559	500	15.36047	0.01	0.9999117
		0.03	0.9998393			0.03	0.9996948
		0.05	0.9996948			0.05	0.9997559
1000	19.00718	0.01	0.9998779	10000	21.48418	0.01	0.999939
		0.03	0.9995079			0.03	0.9999087
		0.05	0.9995422			0.05	0.9997416

By observing Table 2, we can see that when IC ARl is 200, the control limit h of the first stage is 12.20215. On this basis, when FDR is 0.01, the control limit c_h of the second stage is 0.9999559. When FDR is 0.03, the control limit c_h of the second stage is 0.9998393. When FDR is 0.05, the control limit c_h of the second stage is 0.9996948. When the IC ARL is unchanged, c_h decreases as the FDR decreases, which indicates that the control chart can find the OC data streams in time. When the IC ARL decreases, h and c_h both decrease, which indicate that the maximum range of alerts that a control chart can tolerate is increasing. With the increase of c_h, and the monitoring speed increases, the monitoring speed is getting faster and faster, and the effect is getting better and better.

4 CONCLUSIONS

This paper presents a two-stage monitoring procedure for high-dimensional Poisson distribution data. The algorithm satisfies the user's IC ARL and Type-I error rates. In the first stage, the detection delay is greatly reduced by monitoring the global FDR, and the second stage is to quickly find OC data streams by monitoring the local FDR. The algorithm effectively improves the monitoring degree of high-dimensional Poisson distribution data.

In the two-stage monitoring scheme proposed in this paper, the main reason for choosing local FDR as Type-I error rate when identifying OC data streams in the second stage is that global FDR can be easily controlled by controlling local FDR. In some monitoring programs, it may be acceptable to control the Type-I error rate point-by-point when identifying OC data streams. But controlling FDR may be a better choice than controlling the Type-I error rate point-by-point. However, in order to apply the existing FDR control procedure in the second stage, the following *p*-values, denoted by $p_{i,t}^{*}$, essentially need to be calculated based on $G_t > h$,

$$p_{i,t}^{*} = P_{H_{0,i,t}}\left(W_{i,t} > w_{i,t} \mid G_t > h\right),$$

where $w_{i,t}$ is the observed values of $W_{i,t}$.

The calculation of this conditional probability is too cumbersome, so the FDR control procedure cannot be directly applied in the second stage. In future studies, the FDR control procedure proposed by (Storey, 2002)[8] can be applied to the second stage of monitoring.

REFERENCES

[1] Benjamini, Yoav & Hochberg, Yosef. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Journal of the Royal Statistical Society. Series A, 57(1): 289–300.

[2] W. D. EWAN & K. W. KEMP. (1960). Sampling Inspection of Continuous Processes with No Autocorrelation Between Successive Results. J. Biometrika, 47(3): 363–380.

[3] Grigg, O. A. & Spiegelhalter, D. J. (2008). An Empirical Approximation to the Null Unbounded Steady-State Distribution of the Cumulative Sum Statistic. J. Technometrics, 50(4): 501-510.

[4] Jun Li. (2019). A two-stage online monitoring procedure for high-dimensional data streams. J. Journal of Quality Technology, 51(4): 392-406.

[5] Yanting Li & Fugee Tsung. (2009). False Discovery Rate-Adjusted Charting Schemes for Multistage Process Monitoring and Fault Identification. J. Technometrics, 51(2) : 186-205.

[6] Rossi, G. & Lampugnani, L. & Marchi M. (1999). An approximate CUSUM procedure for surveillance of health events. J. Statistics in medicine, 18(16): 2111-22.

[7] Sehgal, V.K. & Kapur, R. & Yadav, K. & Kumar, D. (2010). SOFTWARE RELIABILITY GROWTH MODELS INCORPORATING CHANGE POINT WITH IMPERFECT FAULT REMOVAL AND ERROR GENERATION. J. International Journal of Modelling & Simulation, 30(2): 498-516.

[8] Storey, J. D. (2002). A direct approach to false discovery rate. J. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3): 479–98.

[9] Suvimol, P. & Chananet, C. (2022). Moving Average Control Chart for Generalized Poisson Distribution. J. Journal of Physics: Conference Series, 2346(1): 210-216.

[10] Alevizakos, Vasileios & Chatterjee, Kashinath & Koukouvinos, Christos. (2021). The triple exponentially weighted moving average control chart for monitoring Poisson processes. J. Quality and

Reliability Engineering International, 38(1): 532-549.
[11] Xiao, Yizhuo & Zi, Xuemin. (2021). An MEWMA control chart based on multivariate Poisson distribution data. J. Journal of Physics: Conference Series, 1955(1): 348-353.
[12] Zou, Changliang & Wang, Zhaojun & Zi, Xuemin & Jiang, Wei. (2015). An Efficient Online Monitoring Method for High-Dimensional Data Streams. J. Technometrics, 57(3): 374-87.