Monitoring Data Quality Based on Conditional False Discovery Rate

Abstract

In this paper, we develop a new multivariate exponentially weighted moving average (MEWMA) scheme to on-line monitor the data quality of high-dimensional data streams. With thousands of input data streams, typical metrics like the probability of a false alarm and average run length (ARL) lose their usefulness because one will likely have out-of-control (OC) signals at each time period. Hence, we via a bootstrap approach to control the conditional false discovery rate (CFDR) at each time point given that there is no alarm among all the ongoing streams before the current time point. Numerical results show that the performance of the proposed MEWMA scheme is quite satisfactory.

Keywords: Bootstrap; Data quality; False discovery rate; MEWMA; Statistical process control.

1 Introduction

It has been noted that in today's information age, poor data quality has far-reaching effects and consequences (Redman (1998)). The management of data quality and the quality of associated data management processes have been identified as a critical issue for organizations (e.g., Ballou et al. (1998), Parssian et al. (2004)). There has been extensive research, which has produced a large body of data quality knowledge, and which has played a critical role in our data-intensive, knowledge-based economy (Madnick et al. (2009)). Many of the concepts and procedures of product quality control have been applied to the problem of producing better quality information outputs. Nevertheless, there has been only limited and rudimentary usage of statistical process control (SPC) methods to monitor and improve data quality (Jones-Farmer et al. (2013)). A nice review can be found in Jones-Farmer et al. (2013) and the references therein.

Among the few relevant research on this topic, none used multivariate methods to control the multidimensional data quality process, but instead relied on multiple univariate control
charts. Pierchala et al. (2009) applied 7,569 charts simultaneously to monitor the data quality in the Fatality Analysis Reporting System (FARS). Jones-Farmer et al. (2013) constructed thirty-four similar Bernoulli cumulative sum (CUSUM) charts (Reynolds Jr and Stoumbos (1999)) to monitor the data quality for the cargo aircraft maintenance database. Whenever one wants to monitor several quality variables, there are at least two reasons for introducing a multivariate control procedure. On the one hand, multiple univariate control charts may make determining the control limits by simulation quite complicated (Li et al. (2013)). On the other hand, multivariate methods can take advantage of relationships among quality variables (Woodall and Montgomery (2014)).

In recent years, as discussed by Megahed and Jones-Farmer (2013), the amount of data available for process monitoring continues to grow. The number of variables available in many process-monitoring applications, such as for computer network, healthcare and social networks, has grown tremendously. For an excellent review on the latest developments in this field, we refer to Woodall and Montgomery (2014). Such data streams are sometimes referred to as being “high dimensional”. With thousands of input data streams, typical metrics like the probability of a false alarm and average run length (ARL) lose their usefulness because one will likely have out-of-control (OC) signals at each time period for which data are collected (Woodall and Montgomery (2014)). Thus, under a false discovery rate (FDR) approach, some recent articles control the percentage of signals that are false alarms. However, the traditional techniques and arguments for controlling the FDR in a non-sequential context are not readily extended to the sequential case because at each time point we need to control the FDR given that there is no alarm among all the ongoing streams before the current time point (Du and Zou (2014)). To this end, Du and Zou (2014) propose a procedure which is able to pointwisely control the conditional false discovery rate (CFDR) in time across multiple normal units. In such situations, we are interested in how to monitor the quality of the data itself. In this article, we focus on sequential monitoring the data quality of high-dimensional independent data streams.

With high-dimensional data streams, we observe \( m \) data qualities, such as accuracy, consistency, completeness, etc., for the \( i \)th data stream over time \( t \). If an alarm with respective to a stream is made at the current time, then the monitoring for this stream stopped provisionally. The monitoring of this stream may start over after appropriate adjustment has been made so that the process is in-control (IC) again. Following Jones-Farmer et al. (2013), we suppose that each data quality to be a Bernoulli process. Szarka and Woodall (2011) gave a review on surveillance of Bernoulli processes. To take advantage of the correlation among the data qualities of the same data stream, we construct a multivariate exponentially weighted moving average (MEWMA) chart for each data stream. Here, the null distributions of the MEWMA statistics in different streams may be different. The MEWMA charts (e.g., Joner et al. (2008), Lowry et al. (1992), Zou et al. (2012)) have become as very popular tools in SPC. Runger and Prabhu (1996) developed a Markov Chain method to compute the ARL of MEWMA chart when the observations have a multivariate normal distribution.
Nevertheless, we don’t know the distributions of MEWMA statistics for each data stream, and we can’t apply the Markov Chain method either. Fortunately, we can use Monte Carlo simulation techniques to get the empirical null distributions of MEWMA statistics. Then, we can convert control chart statistics to p-values and employ the FDR based multiple testing procedure of Benjamini and Hochberg (1995) to on-line monitor. More sophisticated alternative multiple testing procedure (e.g., Gavrilov et al. (2009), Storey et al. (2004)) can be used in conjunction with our approach, but that is not the focus of this article. Here, in the sequential case, we focus on how to make use of the information that there is no alarm among all the ongoing streams before the current time point. Hence, we use the conditional null distribution rather than the unconditional null distribution to transform the MEWMA statistic to its p-value. Henceforth, the proposed control scheme using the conditional null distribution is simply called MEWMA scheme, while the scheme using the unconditional null distribution is abbreviated as Uncon scheme. We use a parametric bootstrap approach iteratively to update the empirical conditional null distributions. To decrease the computational load, after the MEWMA sequence of a data stream attains the steady state, we can stop updating the null distribution. By our simulation results, we find that, compared with the Uncon scheme, the proposed MEWMA scheme has more satisfactory empirical FDR and power.

The rest of this paper is organized as follows. In the next section, the statistical model and the MEWMA scheme are introduced. The following section is devoted to investigating the numerical performance of the proposed MEWMA scheme.

2 Methodology

2.1 Model and assumptions

We monitor a large number of $N$ independent data streams over time $t$. For the $i$th data stream, we observe $m_i$ data qualities $X_{i,t} = (X_{i,t,1}, \ldots, X_{i,t,m_i})^T$ such as accuracy, consistency, completeness, etc. For example, let $X_{i,t,j}$ denotes whether the data is accuracy, then

$$X_{i,t,j} = \begin{cases} 0, & \text{if data is accurate,} \\ 1, & \text{if data is inaccurate.} \end{cases}$$

Let $\Sigma_i$ be the phi coefficient (Jones-Farmer et al. (2013)) matrix to assess the degree of correlation among the variables in $X_{i,t}$.

For each stream $i$, $X_{i,1}, X_{i,2}, \ldots$ are supposed to be independent in time domain ($t$). Assume further that under the null hypothesis (IC)

$$Pr\{X_{i,t,j} = 1\} = \mu_{i,j}^0, j = 1, \ldots, m_i,$$
for \( i = 1, \ldots, N \) and \( t = 1, \ldots. \) Under the alternative hypothesis (OC), certain data streams occur changes at some unknown change-points. That is, there exists a sequence of unknown change-points, i.e. \( \tau = (\tau_1, \tau_2, \ldots, \tau_N)^T \), such that the \( i \)th stream is in control with 
\[
Pr\{X_{i,t,j} = 1\} = \mu_{i,j}^0, \ j = 1, \ldots, m_i \text{ if } t \leq \tau_i
\]
and out of control with 
\[
Pr\{X_{i,t,j} = 1\} = \mu_{i,j}^1, \ j = 1, \ldots, m_i \text{ if } t > \tau_i,
\]
for \( i = 1, \ldots, N. \)

In practice, we may be only interested in the case \( \mu_{i,j}^1 > \mu_{i,j}^0 \). The methods discussed below are applicable when \( \tau_i \)'s are different and \( X_{i,t,j}, j = 1, \ldots, m_i \) occur changes at different time.

### 2.2 The MEWMA scheme

It is intended for detecting an increase case \( \mu_{i,j}^1 > \mu_{i,j}^0 \) that we consider a one-sided MEWMA statistic. Alternative one-sided MEWMA approaches can be found in Alessandro (1999), Joner et al. (2008) and Yahav and Shmueli (2013). Here, for relatively simple, we followed Joner et al. (2008). We first express the following EWMA statistics for the \( i \)th data stream at time \( t \)
\[
S_{i,t,j} = \max\{0, (1 - \theta)S_{i,t-1,j} + \theta - X_{i,t,j} - \mu_{i,j}^0 \sqrt{\mu_{i,j}^0(1 - \mu_{i,j}^0)}\}, j = 1, \ldots, m_i.
\]
(2.1)

where \( S_{i,0,j} = 0 \) and \( \theta \in (0, 1) \) is a smoothing parameter. A reflecting boundary (Gan (1993)) is used in equation (2.1) to avoid the inertia problem (Yashchin (1987)). Then, we obtain the MEWMA monitoring statistic
\[
W_{i,t} = S_{i,t}^T \Sigma_i^{-1} S_{i,t},
\]
(2.2)
where \( S_{i,t} = (S_{i,t,1}, \ldots, S_{i,t,m_i})^T, i = 1, \ldots, N \) and \( t = 1, \ldots. \)

Because standardized \( X_{i,t,j} \) is used in equation (2.1), the phi coefficient matrix \( \Sigma_i \) is used directly in equation (2.2). Compared with Joner et al. (2008), we ignore one term including \( \theta \) in equation (2.2) because which can be considered as constant when we use bootstrap approach to compute \( p \)-values. Since no element of \( S_{i,t} \) will ever be less than zero, the resulting chart will signal only as a result of increases in the \( \mu_{i,j} = Pr\{X_{i,t,j} = 1\} \).

We employ the multiple testing procedure of Benjamini and Hochberg (1995) to on-line monitor. The pioneering work of Benjamini and Hochberg (1995) aimed at controlling the FDR instead of the type I error rate at a prespecified level \( \alpha \), while maximizing the number of rejected hypotheses. Let \( H_i^0, i = 1, \ldots, N \), denote
\[
H_i^0: \text{the } i \text{th data stream is IC at current time } t,
\]
and $P_{1,t}, \ldots, P_{N,t}$, the corresponding $p$-values. We use the following Algorithm 1 to on-line monitor the data qualities of $N$ data streams, and use Algorithm 2 to update the empirical $p$-values.

Algorithm 1.
Step 1. Calculate the empirical $p$-values $P_{1,t}, \ldots, P_{N,t}$ by Algorithm 2.
Step 2. Order the empirical $p$-values as $P_{(1),t} \leq \ldots \leq P_{(N),t}$, where $P_{(i),t}$ corresponds to $H^0_{(i)}$.
Step 3. For a pre-chosen level $\alpha$, let $l$ be the largest $i$ for which $P_{(i),t} \leq i\alpha/N$.
Step 4. Find the data-driven threshold $q_t$ at time $t$ by $q_t = P_{(l),t}$.
Step 5. For each stream $i$, if $P_{i,t} \leq q_t$, then the $i$th data stream is halted provisionally, and update $N$.

To clearly explain the algorithm 2, we first consider the monitoring at the time point $t = 1$. For the $i$th data stream, according to the IC $P_{0,j}, j = 1, \ldots, m_i$ and the phi coefficient matrix $\Sigma_i$, we are able to randomly generate pseudo observations $X_{i,1,j,k}$, where $k = 1, \ldots, M$ and $M$ is a sufficiently large integer. Then, we can calculate $M$ observations $S_{i,1,j,k}$ and $W_{i,1,k}$ by equations (2.1) and (2.2) respectively. Next, we get the empirical $p$-values as follow

$$P_{i,1} = \#\{W_{i,1,k} \geq W_{i,1}\}/M,$$

where $i = 1, \ldots, N$.

Then, we consider the monitoring at the time point $t = 2$. Let $P_{i,1} > q_1$, we calculate the $1-q_1$ quantile of $M$ observations $W_{i,1,k}, k = 1, \ldots, M$, and denote it by $W_{i,1-q_1}$. We think the observations $X_{i,1,j,k}, j = 1, \ldots, m_i$ should satisfy the corresponding $W_{i,1,k} < W_{i,1-q_1}$ because the $i$th data stream does not signal alarm when $t = 1$. Thus, we choose all $S_{i,1,j,k}$ satisfied the corresponding $W_{i,1,k} < W_{i,1-q_1}$ as the space of feasible values and randomly bootstrap $M$ observations from the space to make up and update $S_{i,1,j,k}$. The bootstrap approach has been widely applied in SPC. See, for example, Gandy and Kvaløy (2013), Liu and Tang (1996) and Shen et al.(2014). Next, we generate $M$ observations $X_{i,2,j,k}, k = 1, \ldots, M$, and calculate $S_{i,2,j,k}$, $W_{i,2,k}$ by equations (2.1) and (2.2) respectively. The $p$-values can be calculated by

$$P_{i,2} = \#\{W_{i,2,k} \geq W_{i,2}\}/M.$$

The proposed algorithm 2 is summarized as follows.

Algorithm 2.

When $t = 1$, begin with Step 3; otherwise begin with Step 1.
Step 1. Calculate the $1-q_{t-1}$ quantile $W_{i,1-q_{t-1}}$ of $M$ observations $W_{i,t-1,k}, k = 1, \ldots, M$.
Step 2. Bootstrap $M$ observations $S_{i,t-1,j,k}$ satisfied the corresponding $W_{i,t-1,k} < W_{i,1-q_{t-1}}$.
Step 3. Generate $M$ pseudo IC observations $X_{i,t,j,k}, k = 1, \ldots, M$.
Step 4. Calculate $S_{i,t,j,k}$, $W_{i,t,k}$ by equations (2.1) and (2.2) respectively.
Step 5. Calculate the empirical $p$-values by $P_{i,t} = \#\{W_{i,t,k} \geq W_{i,t}\}/M, i = 1, \ldots, N$.

Obviously, Step 2 takes advantage of the useful information that the $i$th data stream is IC at time $t - 1$. Since Algorithm 2 uses the conditional null distribution to transform the
MEWMA statistic to its \( p \)-value, the aim of our proposed MEWMA scheme is to control the CFDR at a prespecified level \( \alpha \), while maximizing the number of rejected hypotheses. Variant concept of CFDR can be found in Du and Zou (2014) and Efron (2007). To use the unconditional null distribution to calculate the \( p \)-values, we can start Algorithm 2 from Step 3. To decrease the computational load, we can stop updating the null distribution (Step 1-4) when \( t \) is sufficiently large.

3 Performance comparisons

In this section, the performance of the proposed MEWMA scheme is compared with the Uncon scheme through Monte Carlo simulations. Each simulation involved 2,000 repetitions. The empirical FDR and empirical power are two criteria used for the performance comparison. Followed Du and Zou (2014), the empirical power at the time point \( t \) is defined as the average proportion of false null hypotheses that are rejected up to time \( t \).

We use the method proposed by Qaqish (2003) to simulate correlated binary variables with specified marginal means and correlations. For completeness, we report the method here. Let \( Y \) denote an \( n \times 1 \) vector of Bernoulli random variables \((Y_1, ..., Y_n)'\), with \( E(Y) = (\mu_1, ..., \mu_n)' = \mu \) and \( corr(Y) = \{r_{jk}\} = R \). We consider the case of equal means and AR(1) correlation for simpleness, say \( \mu_1 = ... = \mu_n, \ r_{jk} = \rho^{|j-k|}, \) for \( j \neq k \) and \(|\rho| < 1 \). We can generate \( y_1, ..., y_n \) sequentially such that \( y_j \sim B(1, \lambda_j), j = 1, ..., n, \) where \( \lambda_1 = \mu_1, \lambda_k = \mu_1 + \rho(y_{k-1} - \mu_1), k = 2, ..., n. \)

For illustrative purpose, we on-line monitor the data qualities of \( N = 1000 \) streams over a period of \( T = 300 \) time-points, where the phi coefficient matrix \( \Sigma_i \) follows AR(1) correlation, and \( m_i = 3, \mu^0_{ij} = 0.05, \mu^1_{ij} = 0.12, i = 1, ..., N, j = 1, ..., m_i. \) We only consider \( \rho = 0.2 \) and \( \rho = 0.5 \) respectively. For each run, 80\% of those streams are IC (\( \tau_i = 300 \)), only other 20\% of the change-points \( \tau_i \sim \text{Poisson}(10) \). For the proposed MEWMA scheme and the Uncon scheme, the parameters \( \theta, M, \alpha \) are chosen as 0.05, 10,000, 0.05 respectively.

The comparisons of empirical power and FDR are shown in Figure 1 and 2 respectively. A desirable method should generally control the FDR below the significance level 0.05, and requires both a much less conservative FDR and high average power. It is obvious that the proposed MEWMA scheme performs much better.

As mentioned before, to decrease the computational load, we can stop updating the null distribution when \( t \) is sufficiently large. We denote the corresponding scheme by MEWMA*. Figure 3(a) compares the empirical power curves of the MEWMA and MEWMA* schemes versus time \([0,300]\). Where the MEWMA* scheme stop updating the null distribution when \( t = 270 \). Hence, we zoom into the corresponding curves of Figure 3(a) versus time \([250,300]\), and display them in Figure 3(b). Figure 3 reveals that the performance of MEWMA and MEWMA* schemes does not have significant difference.
Figure 1: (a)-(b): Empirical power curves versus t.

Figure 2: (a)-(b): Empirical FDR versus t.
4 Concluding Remarks

In this paper, we focus on how control charts may be used to monitor the data quality since the use of SPC methods to monitor and improve the quality of manufacturing and service processes is well researched and implemented in practice. We develop a new false discovery rate-adjusted MEWMA scheme via a bootstrap approach to on-line monitor the data quality of high-dimensional data streams. With thousands of input data streams, we observe $m_i$ data qualities for the $i$th data stream over time $t$. Considering the correlation among the data qualities of the same data stream, we construct a MEWMA chart for each data stream. Supposing different data streams are independent, we use CFDR to obtain the monitoring threshold because it is useless that we optimize ARL when there so many data streams.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China Grants 11071128, 11131002, 11001138, 11101306, 11201246 and the RFDP of China Grant 20110031110002.
References


