Journal of Statistical Computation & Simulation
Vol. 00, No. 00, January 2008, 1–14

RESEARCH ARTICLE

On-line Monitoring Data Quality of High-dimensional Data Streams

Dequan Qi[a,c], Zhonghua Li[b] & Zhaojun Wang[b*]

[a]*LPMC and School of Mathematical Sciences, Nankai University, Tianjin 300071, China*
[b]*LPMC and Institute of Statistics, Nankai University, Tianjin 300071, China*
[c]*Department of Mathematics, Jilin Medical University, Jilin 132013, China*

(*v4.2 released May 2008*)

In recent years, effective monitoring of data quality has increasingly attracted attention of researchers in the area of statistical process control (SPC). Among the relevant research on this topic, none used multivariate methods to control the multidimensional data quality process, but instead relied on multiple univariate control charts. Based on a novel one-sided multivariate exponentially weighted moving average (MEWMA) chart, we propose a conditional false discovery rate-adjusted scheme to on-line monitor the data quality of high-dimensional data streams. With thousands of input data streams, the average run length (ARL) loses its usefulness because one will likely have out-of-control (OC) signals at each time period. Hence, we first control the percentage of signals that are false alarms. Then, we compare the power of the proposed MEWMA scheme with that of two alternative methods. Compared with two competitors, numerical results show that the proposed MEWMA scheme has higher average power.

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

AMS Subject Classification: 62P30; 65C20; 65C60; 68U20

1. Introduction

It has been noted that in today's information age, poor data quality has far-reaching effects and consequences. These impacts include customer dissatisfaction, increased operational cost, less effective decision-making, and a reduced ability to make and execute strategy [1]. The management of data quality and the quality of associated data management processes have been identified as a critical issue for organizations [2, 3]. Therefore, extensive research in the literature has been done, which produced a large body of data quality knowledge, and which played a critical role in our data-intensive, knowledge-based economy. Madnick et al. [4] introduced a framework to characterize the research along two dimensions: topics and methods. Data quality research topics include data quality impact, database-related technical solutions for data quality, data quality in the context of computer science and information technology, and data quality in curation. Data quality research methods include fourteen categories such as action research, artificial intelligence, case study, data mining, design science, mathematical modeling, statistical analysis, etc.

Several researchers have suggested that, like a physical product, data are the end-result of a manufacturing process, with raw data as the input, and a polished,

*Corresponding author. Email: zjwang@nankai.edu.cn

ISSN: 0094-9655 print/ISSN 1563-5163 online

DOI: 10.1080/0094965YYxxxxxxxx
http://www.informaworld.com

transformed data product as the output [5]. Many of the concepts and procedures of product quality control should have been applied to the problem of producing better quality information outputs. The applying control chart methods to enhance data quality is limited and rudimentary. These applications included univariate Shewhart chart, univariate cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) methods. Nevertheless, none used multivariate methods to control the multidimensional data quality process, but instead relied on multiple univariate control charts. A nice review, including the defining, measuring and monitoring data quality, can be found in Jones-Farmer et al. [5] and the references therein. Pierchala et al. [6] applied 7,569 charts simultaneously to monitor the data quality in the Fatality Analysis Reporting System (FARS). Jones-Farmer et al. [5] stated that the false alarm rate would be notably high for this application and there would be little power to detect actual process changes. Whenever one wants to monitor several quality variables, there are another 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 [7]. On the other hand, multivariate methods can take advantage of relationships among quality variables [8].

In recent years, the problem of on-line monitoring a large number of data streams through sequential observations has become increasingly important [9, 10]. As discussed by Woodall and Montgomery [8], the number of variables available in many process-monitoring applications, such as for computer network, healthcare and social networks, has grown tremendously. 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. Thus, under a false discovery rate (FDR) approach, some recent articles control the percentage of signals that are false alarms [11–13]. However, different from the non-sequential context, in the sequential case one should control the FDR given that there is no alarm among all the ongoing streams before the current time point. To this end, Du et al. [14] propose a procedure which is able to control the conditional FDR (CFDR) at each time point. In such situations, we are interested in how to monitor the quality of the data itself via CFDR.

In this paper, we are concerned with sequential monitoring data qualities, such as accuracy, consistency, completeness, etc., of data streams in a situation where the number of data streams is very large. If we suspect data qualities of a data stream deteriorate, then we stop monitoring the corresponding stream provisionally. To monitor these data qualities of high-dimensional data streams, we suggest using one-sided multivariate exponentially weighted moving average (MEWMA) scheme. To make use of the fact that there is no alarm among all the ongoing streams before the current time point, we use the conditional null distribution rather than the unconditional null distribution to transform the MEWMA statistic to its p-value. We propose a novel algorithm to control the CFDR of data streams pointwise in time. We choose the FDR and power as two criteria used for the performance comparison because the ARL loses its usefulness. Numerical results show that the proposed MEWMA scheme has both less conservative FDR and higher average power.

Our contributions are to focus on data qualities of high-dimensional data streams, to use a multivariate method to control the multidimensional data quality process, and to adjust the MEWMA scheme via CFDR to enable it to have less conservative FDR and high average power. The rest of this paper is organized as follows. In the

next section, the statistical model and existing works are presented to illustrate our motivation. Then the proposed one-sided MEWMA scheme is introduced. The following section is devoted to investigating the numerical performance of the proposed MEWMA scheme. Finally, an illustrative example and our conclusions are given.

2. The statistical model and existing works

Both Wang and Strong [15] and Lee et al. [16] organized data quality dimensions into intrinsic and contextual categories. Intrinsic refers to data qualities that are objective and native to the data, such as accuracy, consistency, completeness, etc. Contextual refers to data qualities that are dependent on the context in which the data is observed or used, such as relevancy, believability, accessibility etc. Following Jones-Farmer et al. [5], we limit our study to consider the more general intrinsic measures of data quality, and suppose that each data quality variable can be represented by a Bernoulli process. Topalidou and Psarakis [17] gave a good review of multinomial and multiattribute control charts, including Bernoulli process. The work is, however, not applicable for large number of data streams.

2.1. *The statistical model*

Suppose that there is a large number of independent data streams over time, say, observation $X_{n,t}$ at the nth data stream over time $t = 1, 2, \ldots$ for $n = 1, \ldots, N_t$ with large N_t. Here, we monitor the quality of the data itself. For the nth data stream, we observe m_n data qualities $\mathbf{Y}_{n,t} = (Y_{n,1,t}, \ldots, Y_{n,m_n,t})'$ such as accuracy, consistency, completeness, etc. For example, if $Y_{n,k,t}$ denotes whether the data is accurate, then

$$Y_{n,k,t} = \begin{cases} 0, & X_{n,t} \text{ is accurate,} \\ 1, & X_{n,t} \text{ is inaccurate.} \end{cases}$$

Let $\mu_{n,k,t} = Pr\{Y_{n,k,t} = 1\}$, then $E(Y_{n,k,t}) = \mu_{n,k,t}$ and $Var(Y_{n,k,t}) = \mu_{n,k,t}(1 - \mu_{n,k,t})$ for $k = 1, \ldots, m_n$. To assess the degree of correlation among the variables in $\mathbf{Y}_{n,t}$, let Σ_n be the phi coefficient matrix. The phi coefficient (also known as the mean square contingency coefficient) is a special case of Pearson's correlation coefficient for dichotomous variables [5].

In practice, the data consumers may be more interested in the increase of $\mu_{n,k,t}$, which indicates the quality of the data product has deteriorated. Our statistical model suppose $\mu_{n,k,t}$ changes from $\mu^0_{n,k}$ to another unknown value $\mu^1_{n,k} > \mu^0_{n,k}$ immediately after an unknown change-point. We assume that, for each stream n, $\mathbf{Y}_{n,1}, \mathbf{Y}_{n,2}, \ldots$ are independent in time domain (t), then test the following null and alternative hypotheses at time t. Under the null hypothesis all the data steams are in-control (IC), that is

$$H^0_{n,t} : \{\mu_{n,k,1} = \mu_{n,k,2} \ldots = \mu_{n,k,t} = \mu^0_{n,k}, k = 1, \ldots, m_n\},$$

for $n = 1, \ldots, N_t$. Under the alternative hypothesis, certain data streams occur changes at some unknown change-points, that is,

$$H^1_{n,t} : \{\mu_{n,k,1} = \ldots = \mu_{n,k,(\tau_{n,k})} = \mu^0_{n,k}; \mu_{n,k,(\tau_{n,k}+1)} = \ldots = \mu_{n,k,t} = \mu^1_{n,k}, k = 1, \ldots, m_n\},$$

where $\tau_n = (\tau_{n,1}, \tau_{n,2}, \ldots, \tau_{n,m_n})'$ is an unknown change-point vector.

When $t = 1$, N_1 data streams are observed for the information on data quality. As t increases, some other new data streams may enter into the monitoring system if new information are available. Similarly, some existing data streams may get out of the monitoring system if the data qualities of these data streams are suspected to deteriorate, and these streams may also start over again after appropriate adjustment has been made so that the data qualities are IC again. Considering the observation $\mathbf{Y}_{n,t}$ is a m_n-dimensional vector and the number of data streams N_t is very large, we propose using a multivariate control method incorporating the FDR procedure. We briefly review the existing work in the next subsection.

2.2. *The existing work*

The use of multivariate control charts to monitor and improve the quality of manufacturing and service processes is well researched. Multivariate control charts have a long history in statistics, dating back to 1940's. The chi-squared control chart was described by Hotelling [18]. Multivariate cumulative sum (MCUSUM) and MEWMA control charts have been proposed to improve the performance of a simple chi-squared chart [19–22].

One of the most widely used control charts is the MEWMA chart suggested by Lowry et al. [20] as an extension to its univariate counterpart, which in its simplest form (special case) is defined as follows. Suppose that we observe m-dimensional vector $\mathbf{X}_1, \mathbf{X}_2, \ldots$ with IC mean vector μ_0 and covariance matrix Σ. Let

$$\mathbf{Z}_t = (1 - \lambda)\mathbf{Z}_{t-1} + \lambda(\mathbf{X}_t - \mu_0), \tag{1}$$

where $\mathbf{Z}_0$ is the m-dimensional zero vector and $\lambda \in (0, 1)$ is a smoothing parameter. The chart signals if

$$W_t = \mathbf{Z}_t^T \Sigma_{\mathbf{Z}_t}^{-1} \mathbf{Z}_t > H, \tag{2}$$

where $H > 0$ is chosen to achieve a specified IC ARL. It is known that $\mathbf{Z}_t$ has a covariance matrix equal to

$$\Sigma_{\mathbf{Z}_t} = \frac{\lambda[1 - (1 - \lambda)^{2t}]}{2 - \lambda}\Sigma,$$

or, as $t \to \infty$

$$\Sigma_{\mathbf{Z}_t} = \frac{\lambda}{2 - \lambda}\Sigma.$$

To develop a one-sided MEWMA chart, Joner et al. [23] suggested that a reflecting boundary (cf. [24, 25]) is placed on the EWMA vector given in Equation (1),

$$\mathbf{Z}_t = \max\{0, (1 - \lambda)\mathbf{Z}_{t-1} + \lambda(\mathbf{X}_t - \mu_0)\},$$

where the maximum operator refers to an element-wise comparison of the two vectors. This $\mathbf{Z}_t$ is used in Equation (2) to form the one-sided MEWMA statistic. Alternative one-sided MEWMA approaches can be found in Fassò [26] and Yahav and Shmueli [27].

Generally speaking, three methods, Markov chain, integral equation and Monte Carlo simulation, have been widely used for analyzing the performance of control charts. See Li et al. [28] for an overview of the Markov chain and integral equation methods. As mentioned before, the ARL loses its usefulness when monitor high-dimensional data streams. Therefore, we first briefly review the pioneering work of Benjamini and Hochberg [29]. Benjamini and Hochberg [29] aimed at controlling the FDR instead of the type one error rate at a prespecified level α, while maximizing the number of rejected hypotheses. Consider testing N independent null hypotheses $H_1^0, \ldots, H_N^0$ based on the corresponding p-values $P_1, \ldots, P_N$. Let V be the number of true null hypotheses declared significant and let R be the total of null hypotheses declared significant. The FDR is then defined as $E(Q)$, where $Q = V/R$ is defined as the proportion of the rejected null hypotheses which are incorrectly rejected, with the convention $Q = 0$ when $R = 0$. The procedure proposed by Benjamini and Hochberg [29], also called the BH procedure, runs as follows: let $H_{(i)}^0$ correspond to the ordered p-values $P_{(i)}$, and $J = \max\{1 \leq i \leq N : P_{(i)} \leq i\alpha/N\}$. If such a J exists, reject J hypotheses associated with $P_{(1)}, \ldots, P_{(J)}$; otherwise reject none. More sophisticated alternative multiple testing procedure can be found in Storey et al. [30] and Gavrilov et al. [31]. Some recent articles also applied FDR procedure to Shewhart [32], CUSUM [33] and EWMA [34] charts.

Before ending this section, we briefly review the CFDR procedure proposed by Du et al. [14]. The goal of Du et al. [14] is to sequentially detect the signals of multiple streams simultaneously based on the test statistics $\{S_{i,t}, i \in I_t\}$ using CFDR, where I_t denotes the indices of streams proceeding to time t. Given the fact that the event $G_i^{t-1} = \{S_{i,t-1} \leq q_{t-1}, \ldots, S_{i,1} \leq q_1\}$ has occurred, the ith data stream is OC if $S_{i,t} > q_t$, for some threshold q_t to be determined, where $q_{t'}, t' = 1, \ldots, t-1$, denote the thresholds applied to the test statistics $S_{i,t'}$ with respect to the streams proceeding to time t. The conditional false discovery rate (CFDR) is defined as

$$CFDR(q_t; \mathbf{G}_{t-1}) = E\left(\frac{V(q_t; \mathbf{G}_{t-1})}{R(q_t; \mathbf{G}_{t-1}) \vee 1} | \mathbf{G}_{t-1}\right),$$

where $a \vee b = \max\{a, b\}$, $V(q_t; \mathbf{G}_{t-1}), R(q_t; \mathbf{G}_{t-1})$ denote the number of false rejections and the total number of rejections with respective to the threshold q_t given $\mathbf{G}_{t-1} = \bigcap_{i \in I_t} \{G_i^{t-1}\}$.

Du et al. [14] assume that the observation $X_{i,t}$ for the ith data stream at time t comes from univariate normal distribution, and present some mild conditions on the dependence structure of the stream observations under which CFDR can be controlled. Our statistical model differs from that of Du et al. [14]. We focus on monitoring potentially a large number of streams vectors where elements of the vectors are Bernoulli random variables. In our statistical model, a false discovery would naturally be defined as signalling the stream to be out of control when in fact the observations have been in control since the start. The scheme discussed in the next section is applicable when the τ_n's are different and $\mu_{n,k,t}, k = 1, \ldots, m_n$ occur changes at different times.

3. The proposed one-sided MEWMA scheme

We consider the following EWMA statistics for the n^{th} data stream at time t

$$S_{n,k,t} = (1-\lambda)S_{n,k,(t-1)} + \lambda\frac{Y_{n,k,t}-\mu^0_{n,k}}{\sqrt{\mu^0_{n,k}(1-\mu^0_{n,k})}}, \tag{3}$$

where $S_{n,k,0} = 0, k = 1, \ldots, m_n$. We have $E(S_{n,k,t}) = 0$ if the data qualities of the n^{th} data stream are IC up to time t. The two-sided monitoring statistic according to Lowry et al. [20] can be expressed as

$$W_{n,t} = \frac{2-\lambda}{\lambda[1-(1-\lambda)^{2t}]}\mathbf{S}^T_{n,t}\Sigma_n^{-1}\mathbf{S}_{n,t}, \tag{4}$$

where $\mathbf{S}_{n,t} = (S_{n,1,t}, \ldots, S_{n,m_n,t})^T$.

If we only focus on the deterioration of data qualities, according to Joner et al. [23], we can obtain a one-sided MEWMA monitoring statistic by substituting $\mathbf{S}_{n,t} = (S^J_{n,1,t}, \ldots, S^J_{n,m_n,t})^T$ into Equation (4), where

$$S^J_{n,k,t} = \max\{0, (1-\lambda)S^J_{n,k,(t-1)} + \lambda\frac{Y_{n,k,t}-\mu^0_{n,k}}{\sqrt{\mu^0_{n,k}(1-\mu^0_{n,k})}}\}, k = 1, \ldots, m_n.$$

However, when $t > 1$, $S^J_{n,k,(t-1)}$ is not an unbiased estimator of zero any more under the null hypothesis $H^0_{n,t}$. Therefore, we construct a novel one-sided MEWMA statistic using three steps. We first calculate the EWMA statistics $S_{n,k,t}$ by Equation (3). Then, we modify $S_{n,k,t}$ by

$$S^+_{n,k,t} = S_{n,k,t}I(S_{n,k,t} > 0) = \max\{0, (1-\lambda)S_{n,k,(t-1)} + \lambda\frac{Y_{n,k,t}-\mu^0_{n,k}}{\sqrt{\mu^0_{n,k}(1-\mu^0_{n,k})}}\}. \tag{5}$$

Finally, the $\mathbf{S}_{n,t} = (S^+_{n,1,t}, \ldots, S^+_{n,m_n,t})^T$ is used in Equation (4) to obtain the one-sided MEWMA monitoring statistic.

In practice, some improvement efforts may be undertaken to improve data quality. In this case, having a two sided method becomes important. It is direct to use $S_{n,k,t}$ instead of $S^+_{n,k,t}$ to construct our MEWMA statistic. Here, the statistic $S^+_{n,k,t}$ given in Equation (5) has its own merit. To begin with, the resulting chart will signal only as a result of increases in the $\mu_{n,k,t}$ since no element of $\mathbf{S}_{n,t}$ will ever be less than zero. In addition, our proposal for the one-sided MEWMA statistic is different from that of Joner et al. [23] in the way that we take the maximum of zero and $S_{n,k,t}$, and $S_{n,k,t}$ is an unbiased estimator of zero under the null hypothesis $H^0_{n,t}$.

We propose a simulation-based method to control the CFDR at a prespecified level α, while maximizing the number of rejected hypotheses. The proposed method seems somewhat similar to that of Shen et al. [35], where they focus on a single stream of Poisson count data. We first use the Monte Carlo simulation to get the conditional empirical null distributions of MEWMA statistics. Then, we use the conditional null distributions to transform MEWMA statistics to their p-values. Finally, we employ the BH procedure to on-line monitoring. More sophisticated alternative multiple testing procedure can be used in conjunction with our approach,

but that is not the focus of this paper.

To clearly explain the algorithm, we first consider the monitoring at the time point $t = 1$. For the n^{th} data stream, we randomly generate IC observations $Y_{n,k,1,j}, k = 1, \ldots, m_n, j = 1, \ldots, M$, where M is a sufficiently large integer. Calculate observations $S_{n,k,1,j}$, $W_{n,1,j}$ and the empirical p-values

$$P_{n,1} = \#\{W_{n,1,j} \geq W_{n,1}\}/M.$$

Order the empirical p-values as $P_{(1),1} \leq \ldots \leq P_{(N_1),1}$, where $P_{(i),1}$ corresponds to $H^0_{(i),1}$, for $i = 1, \ldots, N_1$. Let $J = \max\{1 \leq i \leq N_1 : P_{(i)} \leq i\alpha/N_1\}$. If such a J exists, let the data-driven threshold $q_1 = P_{(J),1}$, and reject J hypotheses associated with $P_{(1),1}, \ldots, P_{(J),1}$; otherwise let $q_1 = 0$, and reject none. If an alarm with respective to a stream is made at the current time, 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 IC again.

Then, we consider the monitoring at the time point $t = 2$. Without loss of generality, we assume that the n^{th} data stream does not signal alarm when $t = 1$, say $P_{n,1} > q_1$. If $q_1 > 0$, we calculate the $1 - q_1$ quantile of M observations $W_{n,1,j}, j = 1, \ldots, M$, and denote it by $W_{n,1-q_1}$. We choose all $S_{n,k,1,j}$ satisfied the corresponding $W_{n,1,j} < W_{n,1-q_1}$ as the space of feasible values and randomly generated M observations from the space to make up and update $S_{n,k,1,j}$. If $q_1 = 0$, it is not necessary to update $S_{n,k,1,j}$. Next we randomly generate M observations $Y_{n,k,2,j}, k = 1, \ldots, m_n, j = 1, \ldots, M$, and follow the similar procedure when $t = 1$.

The proposed algorithm is summarized as follows. To decrease the computational load, we can stop updating the null distribution (Step 1-4) when t is sufficiently large (e.g., $t > 100$).

Algorithm (when $t = 1$, begin with Step 3).

Step 1. Calculate the $1 - q_{t-1}$ quantile $W_{n,1-q_{t-1}}$ of M observations $W_{n,t-1,j}$.

Step 2. Generate M observations $S_{n,k,t-1,j}$ satisfied the corresponding $W_{n,t-1,j} < W_{n,1-q_{t-1}}$.

Step 3. For $n = 1, \ldots, N_t$, generate IC observations $Y_{n,k,t,j}, k = 1, \ldots, m_n, j = 1, \ldots, M$.

Step 4. Calculate M observations $S_{n,k,t,j}$, $W_{n,t,j}$ by equations (3) and (4) respectively.

Step 5. Calculate the empirical p-values by $P_{n,t} = \#\{W_{n,t,j} \geq W_{n,t}\}/M$.

Step 6. Order the empirical p-values as $P_{(1),t} \leq \ldots \leq P_{(N_t),t}$, where $P_{(i),t}$ corresponds to $H^0_{(i),t}$, for $i = 1, \ldots, N_t$.

Step 7. For a pre-chosen level α, let J be the largest i for which $P_{(i),t} \leq i\alpha/N_t$.

Step 8. Find the data-driven threshold q_t at time t by $q_t = P_{(J),t}$.

Step 9. If $P_{n,t} \leq q_t$, then the n^{th} data stream is halted provisionally, and update N_t.

4. Performance comparisons

In this section, we compare the proposed MEWMA scheme with two alternative methods to demonstrate the effectiveness of our approach. The first one combines the one-sided MEWMA [23] statistic with the Algorithm in Section 3. The second approach monitors the data qualities using the unconditional null distribution to transform our proposed MEWMA statistic to its p-value. The two competitors are abbreviated as Joner and Uncon schemes hereafter. For a relatively fair comparison, we first control the percentage of signals that are false alarms ($\alpha = 0.05$). Then,

we compare the empirical power. A control scheme with higher average power is considered better. Here, the empirical power at time point t is defined as the average proportion of false null hypotheses that are rejected up to time t. A Fortran program is available from the authors upon request.

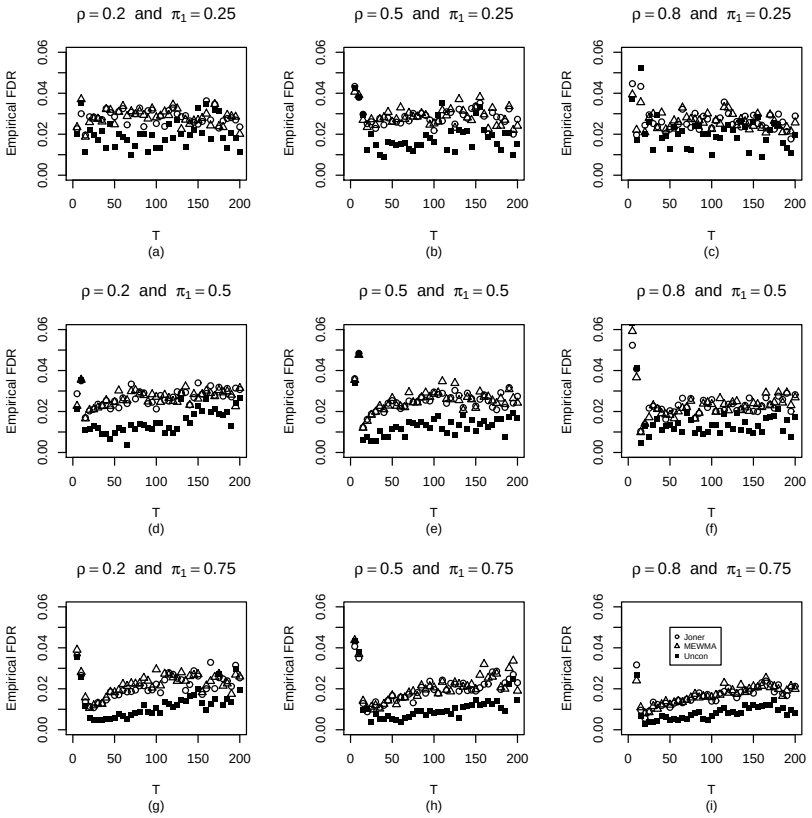

Figure 1. Empirical false discovery rate for the Joner, MEWMA and Uncon control schemes. The legend in the last plot is applicable for all the others.

As mentioned before, we suppose that each data quality to be a Bernoulli process. To simulate correlated binary variables with specified marginal means and correlations, we use the Qaqish [36] method. For completeness, we report the method here. Let $\mathbf{Y}_{n,t}$ denote a $m_n \times 1$ vector of Bernoulli random variables $(Y_{n,1,t}, \ldots, Y_{n,m_n,t})'$, with $E(\mathbf{Y}_{n,t}) = (\mu_{n,1}, ..., \mu_{n,m_n})'$ and $corr(\mathbf{Y}_{n,t}) = \{r_{n,i,j}\} = \Sigma_n$. We consider the case of equal means and AR(1) correlation for simplicity, say $\mu_{n,1} = ... = \mu_{n,m_n}$, $r_{n,i,j} = \rho^{|j-i|}$, for $i \neq j$ and $|\rho| < 1$. According to Qaqish [36], we should generate $y_{n,1,t}, \ldots, y_{n,m_n,t}$ sequentially such that $y_{n,i,t} \sim B(1, \theta_i), i = 1, \ldots, m_n$, where $\theta_1 = \mu_{n,1}, \theta_j = \mu_{n,1} + \rho(y_{n,j-1,t} - \mu_{n,1}), j = 2, \ldots, m_n$. Here, we assume that $m_n = 3$, $\mu^0_{n,k} = 0.05$ and $\mu^1_{n,k} = 0.12$. According to Chaganty and Joe [37], $r_{n,i,j} \geq -0.053$ for $i \neq j$ when the data qualities of the n^{th} stream are IC. Hence, we consider three dependence structures, the values of ρ were chosen as 0.2, 0.5 and 0.8 respectively. Let π_1 denotes the fraction of streams occur changes during the monitoring period $T = 200$. For illustrative purpose, we assume that τ_n's are different but $\mu_{n,k,t}, k = 1, 2, 3$ occur changes at the same time, say $\tau_{n,1} = \tau_{n,2} = \tau_{n,3} \sim$ Poisson(10). Other parameters N_1, λ and M are chosen as 1000, 0.05 and 100,000 respectively, and all the simulation results in this section are based on 1500 replications.

Figure 1 compares the empirical FDR using different control schemes. All schemes generally control the FDR below the significance level 0.05. There is no clear winner

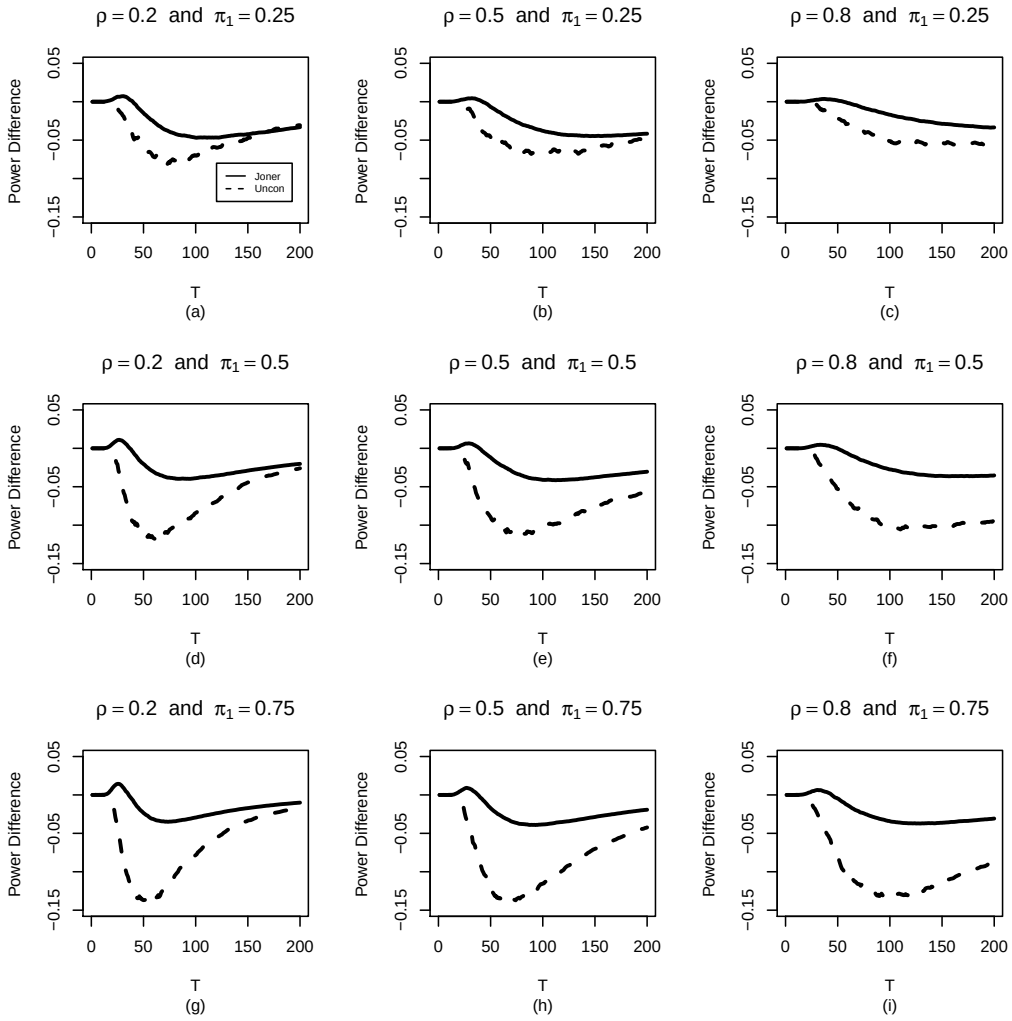

Figure 2. The differences of power from the baseline MEWMA are shown. The legend in the first plot is applicable for all the others.

between the MEWMA and Joner schemes. It confirms that the MEWMA scheme is much less conservative than the Uncon scheme, especially when π_1 is large. Figure 2 illustrates the difference of empirical power from the baseline case (MEWMA scheme), which clearly reveals that the MEWMA scheme performs better than the Uncon scheme, although these benefits become marginal as the time point t increases. Compared with the Joner scheme, the MEWMA scheme performs worse when t is small, but performs better as t increases. For illustrative purpose, some empirical power results of three schemes are presented in Table 1. From Table 1, we can observe that our MEWMA scheme outperforms the Uncon scheme, and the performance of our MEWMA scheme is slightly better than the Joner scheme when $t \geq 50$. Figure 3 displays the conditional cumulative distribution function (CDF) of the proposed MEWMA scheme based on 1 simulation run ($M = 100000$). It is obvious from Figure 3 that the conditional CDFs depend on t, but they are stable when $t \geq 50$. To some extent, this result illustrates that we can stop updating the null distribution when t is sufficiently large as mentioned in Section 3.

Finally, we consider the effect of parameter estimation on our proposed MEWMA scheme. The effect of estimation error has been shown to be a severe problem even when monitoring a single Bernoulli sequence. See Lee et al. [38]. We recommend using the approach proposed by Jones-Farmer et al. [5] to estimate the parameters

Table 1. Comparisons of empirical power

	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.8$		
t	Joner	MEWMA	Uncon	Joner	MEWMA	Uncon	Joner	MEWMA	Uncon
				$\pi_1 = 0.25$					
25	0.051	0.044	0.038	0.031	0.028	0.023	0.016	0.015	0.014
50	0.311	0.326	0.266	0.199	0.206	0.160	0.109	0.107	0.084
100	0.632	0.679	0.609	0.476	0.514	0.447	0.300	0.316	0.265
200	0.868	0.902	0.871	0.749	0.791	0.742	0.548	0.582	0.525
				$\pi_1 = 0.50$					
25	0.091	0.080	0.059	0.052	0.047	0.032	0.023	0.021	0.014
50	0.436	0.457	0.347	0.302	0.314	0.224	0.172	0.172	0.119
100	0.759	0.798	0.714	0.618	0.659	0.555	0.429	0.456	0.355
200	0.932	0.952	0.926	0.853	0.883	0.828	0.694	0.729	0.635
				$\pi_1 = 0.75$					
25	0.133	0.119	0.077	0.076	0.067	0.045	0.034	0.029	0.018
50	0.547	0.570	0.434	0.396	0.413	0.296	0.240	0.245	0.164
100	0.852	0.882	0.803	0.734	0.772	0.657	0.552	0.586	0.455
200	0.970	0.980	0.963	0.924	0.943	0.901	0.815	0.846	0.757

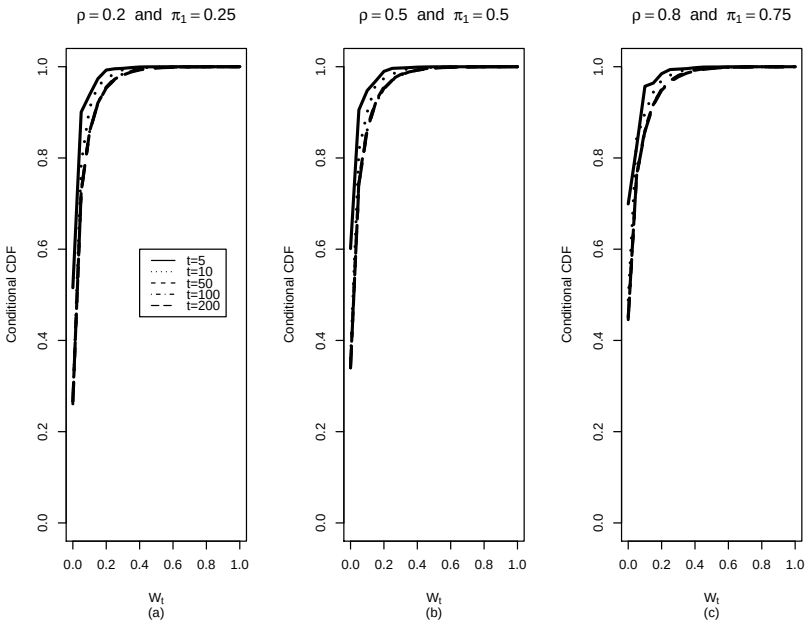

Figure 3. The conditional cumulative distribution function (CDF) of the proposed MEWMA scheme based on 1 simulation run ($M = 100000$).

$\mu^0_{n,k}$ and Σ_n based on Phase I data. For illustration purpose, we randomly generate IC data with sample size m equals 1000, 10000 and 100000, respectively, as Phase I data. Here, $\mu^0_{n,k}$ is estimated by its maximum likelihood estimator, Σ_n is estimated by the Pearson's correlation coefficient matrix. At each time point, we generate $M = 100000$ observations from Phase I data to calculate the empirical p-values used for online monitoring. The empirical FDR and power of the proposed

MEWMA scheme based on different number of Phase I data are shown in Figure 4. From Figure 4, we find that there is no evident difference between the results of $m = 10000$ and $m = 100000$. In addition, apparently, the results of $m = 1000$ seems worse than those of the other two cases. Therefore, in Phase I control, it is important that the sample size m is sufficiently large so that the performance of our proposed MEWMA scheme in Phase II monitoring will be affected negligibly.

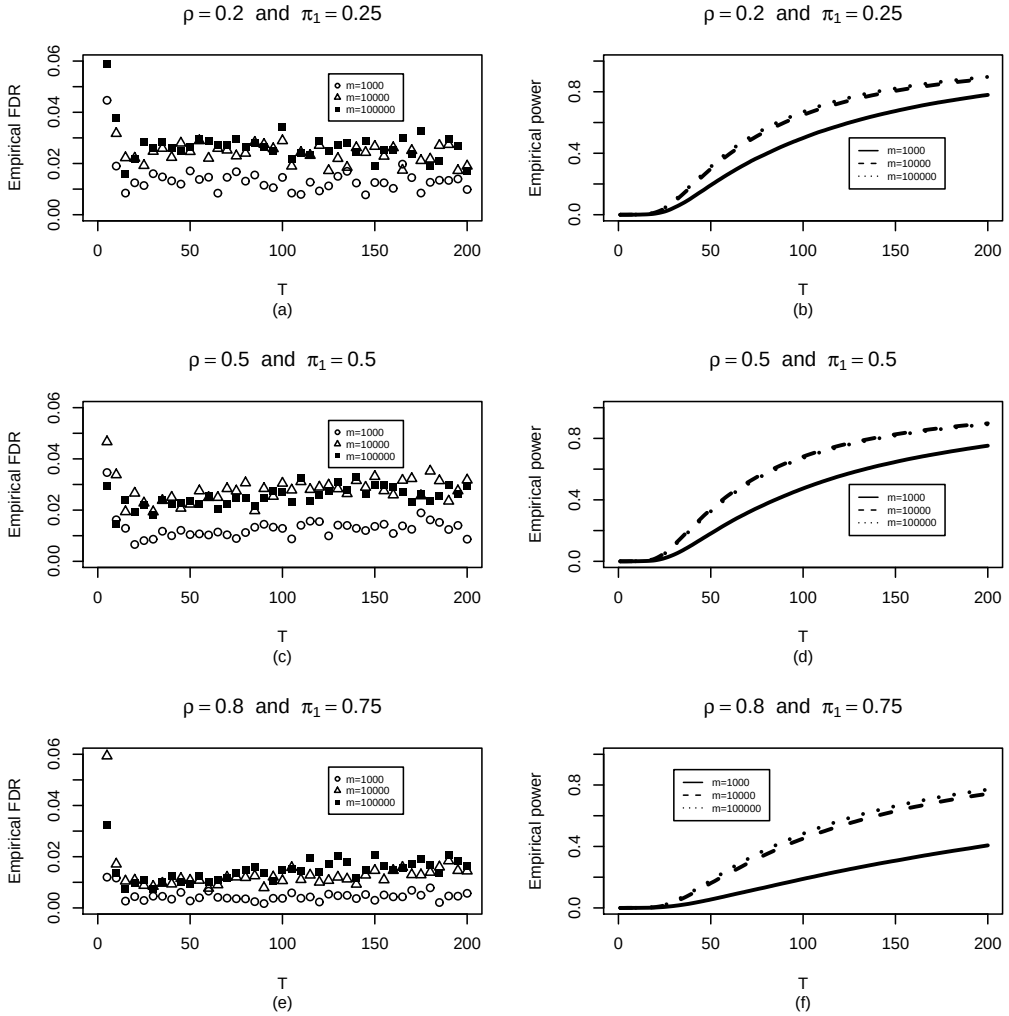

Figure 4. The empirical FDR and power of the proposed MEWMA scheme based on different numbers of Phase I data.

5. Illustrative Example

In this section, we use one example to illustrate how to apply the developed methods in practice. Jones-Farmer et al. (2014) integrated a discussion of a data production process based on the cargo aircraft maintenance example. In the aircraft maintenance database, there are 603 records. Each record has 14 separate data properties which can be measured with quality metrics of accuracy, completeness, and consistency. The data quality measures for this database can be summarized to include 34 dichotomous variables. Jones-Farmer et al. (2014) constructed 34 similar Bernoulli CUSUM charts to monitor the data qualities. Generally, if the above 14 separate data properties can be looked as 14 data streams, the question

is how to monitor the data qualities of several data streams. We choose 10 out of them according to Table 2 and Table 3 in Jones-Farmer et al. (2014), and adjust the phi coefficient. The IC mean and phi coefficient were shown in Table 2.

Table 2. The IC mean and phi coefficient

n	m_n	$\mu^0_{n,1}$	$\mu^0_{n,2}$	$\mu^0_{n,3}$	$r_{n,1,2}$	$r_{n,1,3}$	$r_{n,2,3}$
1	3	0.0006	0.0006	0.0018	0	0	0
2	3	0.3272	0.0006	0.0006	0	0	0
3	3	0.1818	0.0006	0.0006	0	0	0
4	3	0.4790	0.0006	0.0006	0	0	0
5	3	0.2834	0.0006	0.0006	0	0	0
6	3	0.4662	0.0006	0.0006	0	0	0
8	3	0.1993	0.0006	0.3272	0	0.18	0
8	2	0.2687	0.2687	–	-0.2	–	–
9	2	0.1042	0.1042	–	-0.1	–	–
10	2	0.0018	0.0018	–	0.2	–	–

To demonstrate the effectiveness of our proposed method, we assume that the OC mean $\mu^1_{n,k} = \mu^0_{n,k} + \frac{1}{2}\sqrt{\mu^0_{n,k}(1-\mu^0_{n,k})}$. Supposing $N_1 = 1000$, we randomly generate the data qualities of different data streams according to Table 2. We assume that $\pi_1 = 75\%$ of streams occur changes during the monitoring period $T = 100$. Assume further that the change point time τ_n's are different and $\mu_{n,k,t}$ occur changes at the different time, say $\tau_{n,k} \sim$ Poisson(10), $k = 1, \ldots, m_n$. Other parameters λ, α and M are chosen as 0.05, 0.05 and 100,000 respectively. We aim to investigate the number of discoveries at time t. We are also interested in the power and the false discovery proportion up to time t (TFDP). The simulation results are shown in Figure 5 from which we can observe that the proposed procedure works reasonably well.

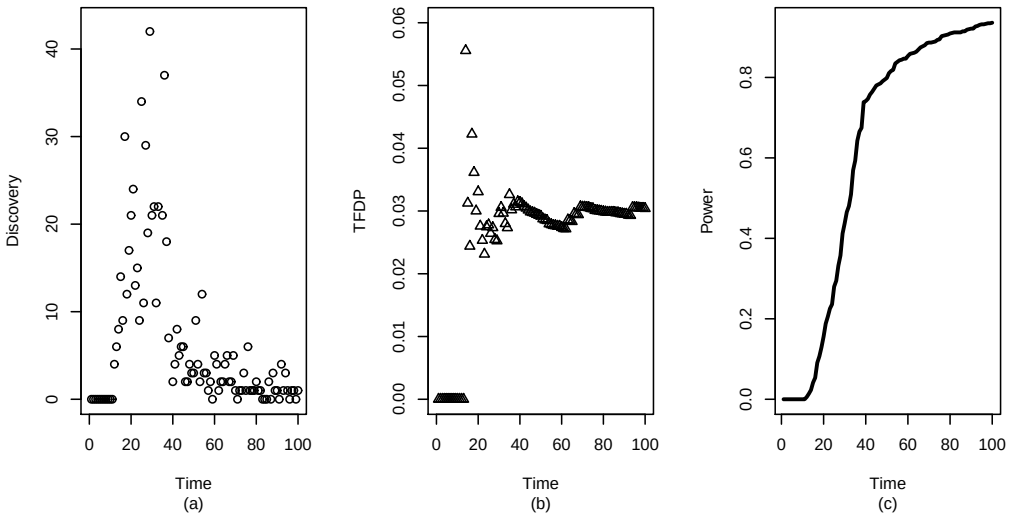

Figure 5. (a)The number of discoveries; (b) TFDP; (c)Power curve.

6. Concluding Remarks

We have presented a new framework to on-line monitor the intrinsic data qualities, such as accuracy, consistency, completeness, etc., of a large number of data streams based on the assumption of independence within data streams. For each data stream, every data quality observation is a multidimensional variable with marginal Bernoulli distribution. By this framework, a novel one-sided multivariate exponentially weighted moving average (MEWMA) statistic is first calculated, and then BH procedure is used to control the conditional false discovery rate (CFDR). Compared to existing statistical process control (SPC) methods that are applied to monitor data quality, the proposed framework has the contributions that (i) it focuses on data qualities of high-dimensional data streams, (ii) it uses multivariate method to control the multidimensional data quality process, (iii) it adjusts the MEWMA scheme via CFDR. It still requires much future research on how to on-line monitor under some special dependence structures and in cases when the marginal Bernoulli distribution with a very low proportion nonconforming.

Acknowledgements

The authors would like to thank the Editor, the Associate Editor and two Referees for their insightful comments. This paper is supported by the National Natural Science Foundation of China Grants 11431006, 11371202, 11131002, 11571191, 11201246 and 11401573.

References

[1] Redman, T. C. The Impact of Poor Data Quality on the Typical Enterprise. Commun. ACM. 1998;41:79–82.
[2] Ballou, D., Wang, R., Pazer, H., Tayi, G. K. Modeling Information Manufacturing Systems to Determine Information Product Quality. Manage. Sci. 1998;44:462–484.
[3] Parssian, A., Sarkar, S., Jacob, V. S. Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. Manage. Sci. 2004;50:967–982.
[4] Madnick, S. E., Wang, R. Y., Lee, Y. W., Zhu, H. Overview and Framework for Data and Information Quality Research. J. Data Inform. Qual. 2009;1:1–22.
[5] Jones-Farmer, L. A., Ezell, J. D., Hazen, B. T. Applying Control Chart Methods to Enhance Data Quality. *Technometrics*. 2014;56:29–41.
[6] Pierchala, C. E., Surti, J., Peytcheva, E., Groves, R. M., Kreuter, F., Kohler, U., Chipperfield, J. O., Steel, D. G., Graham, P., Young, J. Control Charts as a Tool for Data Quality Control. J. Off. Stat. 2009;25:167–191.
[7] Li, J., Tsung, F., Zou, C. Multivariate Binomial/Multinomial Control Chart. IIE Trans. 2014;46:526–542.
[8] Woodall, W. H., Montgomery, D. C. Some Current Directions in the Theory and Application of Statistical Process Monitoring. J. Qual. Technol. 2014;46:79–94.
[9] Mei, Y. Efficient Scalable Schemes for Monitoring a Large Number of Data Streams. Biometrika. 2010;97:419–433.
[10] Zou, C., Jiang, W., Wang, Z., Zi, X. An Efficient On-line Monitoring Method for High-dimensional Data Streams. Technometrics. 2014; DOI:10.1080/00401706.2014.940089, available at http://amstat.tandfonline.com/doi/abs/10.1080/00401706.2014.940089#.VBlSMNqS1dg.
[11] Li, Y., Tsung, F. False Discovery Rate-Adjusted Charting Schemes for Multistage Process Monitoring and Fault Identification. *Technometrics*. 2009;51:186–205.
[12] Spiegelhalter, D., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C., Grigg, O. Statistical Methods for Healthcare Regulation: Rating, Screening and Surveillance (With Discussions). J. R. Statist. Soc. A. 2012;175:1–47.
[13] Gandy, A., Lau, F. D. Non-restarting Cumulative Sum Charts and Control of the False Discovery Rate. Biometrika. 2013;100:261–268.
[14] Du, L., Zhang, C., Zou, C. On-line Control of False Discovery Rates with Application to SPC. Submitted for publication. 2014.
[15] Wang, R. Y., Strong, D. M. Beyond Accuracy: What Data Quality Means to Data Consumers. J. Manage. Inform. Syst. 1996;12:5–33.
[16] Lee, Y. W., Strong, D. M., Kahn, B. K., Wang, R. Y. AIMQ: a Methodology for Information Quality Assessment. Inform. Manage-Amster. 2002;40:133–146.

[17] Topalidou, E., Psarakis, S. Review of Multinomial and Multiattribute Quality Control Charts. Qual. Reliab. Eng. Int. 2009;25:773–804.
[18] Hotelling, H. H. Multivariate Quality Control Illustrated by the Air Testing of Sample Bombsights. in Techniques of Statistical Analysis, eds. C. Eisenhart, M. W. Hastay, and W. A. Wallis, New York: McGraw-Hill. 1947;pp. 111–184.
[19] Crosier, R. B. Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. Technometrics. 1988;30:243–251.
[20] Lowry, C. A., Woodall, W. H., Champ, C. W., Rigdon, S. E. A Multivariate Exponentially Weighted Moving Average Control Chart. Technometrics. 1992;34:46–53.
[21] Qiu, P., Hawkins, D. M. A Nonparametric Multivariate CUSUM Procedure for Detecting Shifts in All Directions. J. R. Statist. Soc. D. 2003;52:151–164.
[22] Zou, C., Wang, Z., Tsung, F. A Spatial Rank-Based Multivariate EWMA Control Chart. Nav. Res. Log. 2012;59:91–110.
[23] Joner, M. D. Jr., Woodall, W. H., Reynolds, M. R. Jr., Fricker, R. D. Jr. A One-Sided MEWMA Chart for Health Surveillance. Qual. Reliab. Eng. Int. 2008;24:503–518.
[24] Gan, F. Exponentially Weighted Moving Average Control Charts with Reflecting Boundaries. J. Stat. Comput. Sim. 1993;46:45–67.
[25] Li, Z., Wang, Z., Wu, Z. Necessary and Sufficient Conditions for Non-interaction of a Pair of One-Sided EWMA Schemes with Reflecting Boundaries. Stat. Probabil. Lett. 2009;79:368–374.
[26] Fassò, A. One-sided MEWMA Control Charts. Commun. Stat.-Simul. C. 1999;28:381–401.
[27] Yahav, I., Shmueli, G. Directionally Sensitive Multivariate Control Charts in Practice: Application to Biosurveillance. Qual. Reliab. Eng. Int. 2014;30:159–179.
[28] Li, Z., Zou, C., Gong, Z., Wang, Z. The Computation of Average Run Length and Average Time to Signal: An Overview. J. Stat. Comput. Sim. 2014;84:1779–1802.
[29] Benjamini, Y., Hochberg, Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. J. R. Statist. Soc. B. 1995;57:289–300.
[30] Storey, J. D., Taylor, J. E., Siegmund, D. Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: a Unified Approach. J. R. Statist. Soc. B. 2004;66:187–205.
[31] Gavrilov, Y., Benjamini, Y., Sarkar, S. K. An Adaptive Step-down Procedure with Proven FDR Control under Independence. Ann. Statist. 2009;37:619–629.
[32] Lee, S. H., Jun, C. H. A Process Monitoring Scheme Controlling False Discovery Rate. Commun. Stat.-Simul. C. 2012;41:1912–1920.
[33] Li, Y., Tsung, F. Multiple Attribute Control Charts with False Discovery Rate Control. Qual. Reliab. Eng. Int. 2012;28:857–871.
[34] Lee, S. H., Park J. H., Jun, C. H. An Exponentially Weighted Moving Average Chart Controlling False Discovery Rate. J. Stat. Comput. Sim. 2014;84:1830–1840.
[35] Shen, X., Zou, C., Jiang, W., Tsung, F. Monitoring Poisson Count Data with Probability Control Limits When Sample Sizes are Time Varying. Nav. Res. Log. 2013; 60:625–636.
[36] Qaqish, B. F. A Family of Multivariate Binary Distributions for Simulating Correlated Binary Variables with Specified Marginal Means and Correlations. Biometrika. 2003;90:455–463.
[37] Chaganty, R., Joe, H. Range of Correlation Matrices for Dependent Bernoulli Random Variables. Biometrika. 2006;93:197–206.
[38] Lee, J., Wang, N., Xu, L., Schuh, A., Woodall, W. H. The Effect of Parameter Estimation on Upper-sided Bernoulli Cumulative Sum Charts. Qual. Reliab. Eng. Int. 2013;29:639–651.