

Article

Not peer-reviewed version

An Online Monitoring and Fault Diagnosis Scheme for High-Dimensional Stream

[Tao Wang](#)^{*}, Yunfei Guo, Fubo Zhu

Posted Date: 21 January 2025

doi: 10.20944/preprints202501.1422.v1

Keywords: High-dimensional data stream; Statistical process control; Online monitoring; Statistical diagnosis



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

An Online Monitoring and Fault Diagnosis Scheme for High-Dimensional Stream

Tao Wang ^{1,*}, Yunfei Guo ² and Fubo Zhu ¹¹ School of Mathematics and Statistics, Huaiyin Normal University, Huai'an 223300, China² Department of Mathematics, Yanbian University, Yanji City 133002, China

* Correspondence: skywangtao2007@126.com

Abstract: This article proposes a two-stage process monitoring and fault diagnosis scheme for the high-dimensional data stream. In the first stage, we choose the exponentially weighted moving average (EWMA) statistic to construct the online monitoring statistic, then determine the change point of the data stream based on extreme value theory and multiple hypothesis testing procedures. In the second stage, when the data stream has an abnormal change alarm, it enters an out-of-control state, and we perform a fault diagnosis procedure on the data stream. Specifically, when an abnormal alarm occurs, we further utilize the numerical simulation procedures to detect these components that exhibit anomalies. In addition, to verify the performance of the proposed method, this work also conducts numerical simulations and case studies and compares them with other competitive methods.

Keywords: high-dimensional data stream; statistical process control; online monitoring; statistical diagnosis

1. Introduction

In recent years, with the rapid development of modern sensing and data acquisition technology, real-time acquisition of high-dimensional data streams has become feasible in many application scenarios. Due to some observations that may shift in the data stream, the process transitions from the in-control (IC) state to an out-of-control (OC) state. Before proceeding, for an intuitive explanation, we present a simple example to illustrate IC and OC observations in a multi-dimensional data stream. As shown in Figure 1, a p -dimensional ($p = 20$) time series data stream over $t = 300$ time points is recorded. The IC observations X_t s are independently and identically distributed (i.i.d.) multivariate normal random variables with the mean vector $\mu_0 = (0, 0, \dots, 0)'_p$, and the covariance matrix $\Sigma_p = I_p$. After a time point $\tau = 200$, p_s components of the high-dimensional data experience an anomaly, resulting in a change of the data stream. This example mentioned two types of abnormal changes in a time series data stream: the mean drift change and the change of the covariance matrix. In this work, we only study data streams with mean drift changes.

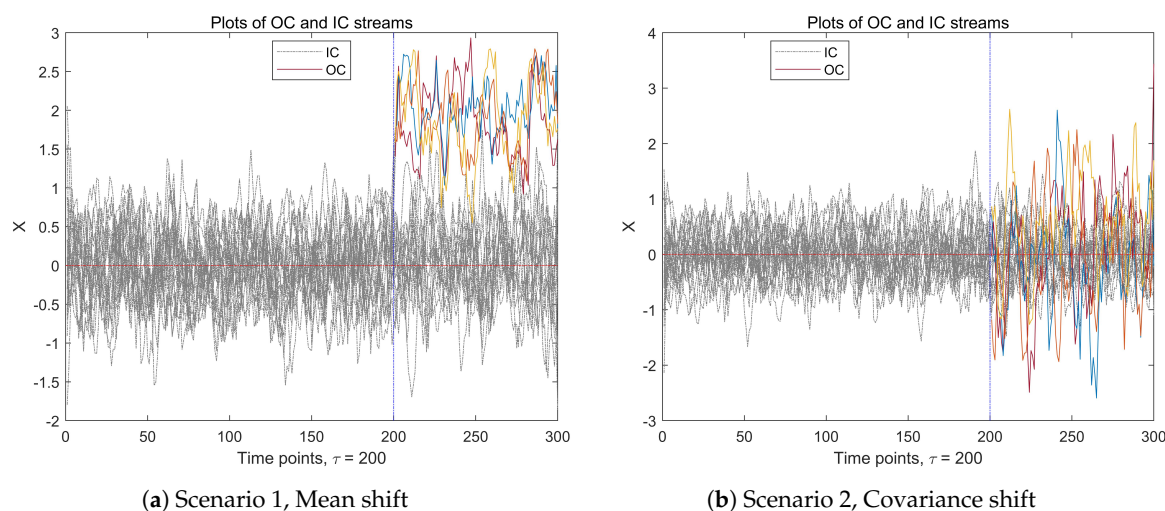


Figure 1. The data stream with IC data (gray) and OC data (color), when $p = 20$, $p_s = 4$, $\tau = 200$, $t = 300$, $\kappa = 2.0$. The left subfigure is mean-shift case, and the right subfigure is the covariance shift setting.

As a result of the fact that the data stream may be high-dimensional and the variables are complexly related, we need to address the following two tasks: The first issue is online monitoring of the high-dimensional data stream, which aims to check whether the data stream has changed over time points and accurately estimates the change point τ after the change occurs. The second issue is fault diagnosis, which aims to identify the components that have undergone abnormal changes, it will help eliminate the root causes of the change.

The problem of process monitoring and diagnosis for high-dimensional streaming data based on statistical process control (SPC) methods has become an active research area. A series of articles has focused on sequential sampling for monitoring and diagnosis of high-dimensional stream. Zou et al.(2015)[1] developed a control chart based on a powerful goodness-of-fit test of the local cumulative sum statistics from each data stream for monitoring high-dimensional data streams. Li (2019)[2] proposed a two-stage monitoring procedure to control both the IC-ARL and Type-I errors at the levels specified by users. Li et al.(2020) [3] provide a computationally fast diagnostic procedure for high-dimensional data streams to control the weighted missed discovery rate at some satisfactory level. Ahmadi-Javid and Ebadi (2021) [4] proposed a two-step method for monitoring high-dimensional normally distributed multi-stream processes. Ebrahimi et al. (2021) [5] proposed an integrated monitoring and diagnostics approach for large-scale streaming data. For more recent research on the theory, method and application of monitoring and diagnosis for high-dimensional data streams, see ref. ([6–14]).

As the EWMA statistic is sensitive to small changes and includes the information of previous samples, thus we use it for online monitoring ([15]). At the first stage, we construct a monitoring statistic based on the EWMA statistic. Under certain conditions, we prove the asymptotic distribution of the monitoring statistic, and then obtain the threshold of online monitoring. At any time point, if the monitoring statistic value is greater than the threshold, a real-time alert is issued, indicating that the data stream is out of control. Hence, we move into the second stage, to identify the components that have undergone abnormal changes over time points, which belongs to the scope of fault diagnosis. Specifically, when the monitoring procedure proposed in the first stage alerts that the data stream is abnormal, we allow the monitoring procedure to continue running for a while, and obtain more sample data, then we conduct multiple hypothesis tests on the data before and after the alarm to determine which components of the data stream have experienced anomalies. The desired two-stage monitoring and diagnosis scheme in this study can not only correctly estimate the time point in time, but also identify the OC components of the data stream accurately.

The remainder of this article is arranged as follows. In Section 2, we formulate the online monitoring and fault diagnosis problem, present the preliminary research plan, and provide online monitoring and diagnosis strategies; furthermore, we provide the evaluation measures for monitoring and diagnostic results of high-dimensional data streams. In Section 3, we conduct numerical simulations on the proposed method and compare it with some other competitive methods.. Section 4 illustrates the practicality of the proposed method through a real data stream. The conclusions of this article are provided in Section 5.

2. Methodology

2.1. Problem Formulation

Suppose that we monitor a high-dimensional data stream $X_1, X_2, \dots, X_t, \dots$ over time points, where $X_t = (x_{1t}, x_{2t}, \dots, x_{pt})' \in \mathbb{R}^p$ denotes a p -dimensional observation at time point t . In the IC state, the observations are generalized from $N(\mu_0, \Sigma_p)$, with $\mu_0 = (\mu_{01}, \dots, \mu_{0p})' \in \mathbb{R}^p$ and $\Sigma_p = (\sigma_{jk})_{p \times p}$, $j, k = 1, \dots, p$; however, after an unknown change point τ , the mean value changes from μ_0 to μ_1 , and the stream incurs mean shift and becomes OC. In many applications, there are usually only a few components of the data stream responsible for process changes, thus the shift $\mu_1 - \mu_0$ is sparse with a small number of non-zero variables, while the majority of variables equal to zero.

As a result, this paper focuses on addressing the following two issues: The first one is to propose an effective online monitoring method that can accurately estimate the location of the change point. Once the data flow changes abnormally, it can issue an alarm in time. The second task is to accurately identify the components of abnormal changes in high-dimensional data streams while controlling the error rate after the alarm is issued.

2.2. The EWMA statistic

As is well known, the EWMA statistic is sensitive to small changes, and it includes information from previous samples [16]. Thus, an EWMA statistic is chosen in this study for monitor high-dimensional data stream, the multivariate EWMA sequence is defined by

$$Y_t = \gamma(X_t - \mu_0) + (1 - \gamma)Y_{t-1}, \quad t = 1, 2, \dots \quad (1)$$

where $\gamma \in [0, 1]$ is a weigh parameter. As described in Feng et al(2020) [17], a smaller value of γ leads to a quicker detection of smaller shifts, the detailed comparison results are provided in the numerical simulation section.

Let $Y_t = (y_{1t}, y_{2t}, \dots, y_{pt})'$, the j -th component of Y_t is denoted as

$$y_{jt} = \gamma(x_{jt} - \mu_{0j}) + (1 - \gamma)y_{j(t-1)}, \quad j = 1, \dots, p, \quad t = 1, 2, \dots, \quad (2)$$

where $Y_0 = (0, 0, \dots, 0)'$ is pre-arranged, i.e., $y_{j0} = 0$ for $j = 1, \dots, p$. It is obvious that when the data stream is IC, we have

$$E(Y_t) = (0, 0, \dots, 0)', \quad Cov(Y_t) = \left(\frac{\gamma}{2 - \gamma}\right)(1 - (1 - \gamma)^{2t})\Sigma,$$

As $t \rightarrow \infty$, we have $(1 - (1 - \gamma)^{2t}) \rightarrow 1$, then the asymptotic distribution of Y_t is as follows,

$$Y_t \sim N_p(\mathbf{0}, \frac{\gamma}{2 - \gamma}\Sigma), \quad t \rightarrow \infty. \quad (3)$$

Correspondingly, the j -th component of Y_t obeys the following normal distribution, i.e.,

$$y_{jt} \sim N(0, \frac{\gamma}{2 - \gamma}\sigma_{jj}), \quad t \rightarrow \infty.$$

where σ_{jj} is the j -th diagonal element of Σ . However, once the mean value has undergone abnormal changes, Y_t does not follow the multivariate normal distribution mentioned above.

2.3. Online Monitoring Procedure

In fact, the online monitoring of high-dimensional data streams can be considered as the following hypothesis testing problem, with the null and the alternative hypothesis:

$$\begin{aligned}\mathcal{H}_{0t} : Y_1, Y_2, \dots, Y_t &\sim N(0, \frac{\gamma}{2-\gamma}\Sigma), \\ \mathcal{H}_{1t} : \exists \tau > 0, Y_1, Y_2, \dots, Y_{\tau-1} &\sim N(0, \frac{\gamma}{2-\gamma}\Sigma), \\ \text{but } Y_\tau, Y_{\tau+1}, \dots, Y_t &\text{ do not follow } N(0, \frac{\gamma}{2-\gamma}\Sigma).\end{aligned}\quad (4)$$

It is obvious that the monitored data stream is IC up to time point t if $\mathcal{H}_{0t}$ is accepted, but a rejection of $\mathcal{H}_{0t}$ indicates that some components of the stream have changed abnormally after the time point τ .

Therefore, we construct the test statistic as the following max-normal form,

$$M_t = \max_{1 \leq j \leq p} \frac{y_{jt}^2}{\frac{\gamma}{2-\gamma}\sigma_{jj}}, \quad t = 1, 2, \dots \quad (5)$$

Intuitively, the larger value of M_t , the more likely the observation undergoes abnormal at time point t , if M_t exceeds a threshold, which implies that an anomaly occurred.

It should be noted that the mean and covariance values of Y_t are unknown, so the reasonable estimators of μ and Σ are needed. Suppose $X_1, X_2, \dots, X_m$ be the initial IC data stream, the sample mean and covariance of these m observations are defined as $\hat{\mu}_0 = \frac{1}{m} \sum_{t=1}^m X_t$ and $\hat{\Sigma} = \frac{1}{m-1} \sum_{t=1}^m (X_t - \hat{\mu}_0)(X_t - \hat{\mu}_0)'$, respectively.

Consequently, the estimated max-normal monitoring statistic is defined as

$$\tilde{M}_t = \max_{1 \leq j \leq p} \frac{y_{jt}^2}{\frac{\gamma}{2-\gamma}\hat{\sigma}_{jj}}, \quad t = m+1, m+2, \dots \quad (6)$$

where $\hat{\sigma}_{jj}$ is the j -th main diagonal element of $\hat{\Sigma}$.

For simplicity, let $z_{jt} = \frac{y_{jt}}{\sqrt{\frac{\gamma}{2-\gamma}\hat{\sigma}_{jj}}}$, $j = 1, \dots, p$, it is obvious that, $z_{jt} \sim N(0, 1)$, $t \rightarrow \infty$. The monitoring statistic can be rewritten as

$$\tilde{M}_t = \max(z_{jt}^2). \quad (7)$$

Let $Z_t = (z_{1t}, z_{2t}, \dots, z_{pt})'$, then Z_t is a high-dimensional normal random vector with the mean vector of zero and the covariance matrix of $\Omega = (\omega_{jk})_{p \times p}$ for $1 \leq j, k \leq p$, where the diagonal $\omega_{jj} = 1$ for $1 \leq j \leq p$. In order to obtain the threshold of online monitoring, we need to prove the asymptotic distribution of the statistic $\tilde{M}_t$ when the null hypothesis holds.

To this end, this requires the following two assumptions:

(A.1) $\max_{1 \leq j \neq k \leq p} |\omega_{i,j}| \leq r < 1$ for some constant $0 < r < 1$;

(A.2) $\max_{1 \leq j \neq k \leq p} \omega_{i,j}^2 \leq C_0$ for some constant C_0 ;

Assumption (A1) and (A2) are mild and common assumptions that are widely used in the high-dimensional means test under dependence. Under the above two assumptions, Proposition 1 studies the asymptotic distribution of $\tilde{M}_t$ when the null hypothesis holds.

Proposition 1. Under the assumptions of (A.1) and (A.2), when the null hypothesis $\mathcal{H}_{0t}$ holds, for any $z \in \mathbb{R}$, $\tilde{M}_t$ follows the the asymptotic distribution,

$$P(\tilde{M}_t - 2\log(p) + \log(\log(p)) \leq z) \rightarrow \exp\left\{-\frac{1}{\sqrt{\pi}}\exp\left\{-\frac{z}{2}\right\}\right\}, \quad \text{as } m, p \rightarrow \infty. \quad (8)$$

Proposition 1 shows that $\tilde{M}_t - 2\log(p) + \log(\log(p))$ follows the type I extreme value distribution with the cumulative distribution function $\exp\{-\frac{1}{\sqrt{\pi}}\exp\{-\frac{z}{2}\}\}$. The proof of this proposition is similar to that of Lemma 6 in Cai et al. 2014 [18], then we omit here.

Thus, the data stream shows abnormal change at time point t , if

$$\tilde{M}_t > L_\alpha, \quad (9)$$

where $L_\alpha = 2\log p - \log(\log p) + q_\alpha$ is the threshold at a given significance level α , and $q_\alpha = -\log(\pi) - 2\log(\log(1 - \alpha)^{-1})$.

It should be noted that not all the observations of $\tilde{M}_t > L_\alpha$ are considered as change points, and these points may also be outliers. Therefore, when $\tilde{M}_t > L_\alpha$ occurs, we continue to monitor for a while to obtain some more observations, and if all the values of $\tilde{M}_t$ are greater than L_α . Thus, the change-point τ can be estimated naturally by

$$\hat{\tau} = \inf\{t : t > t_{ic} | \tilde{M}_{t+i} > L_\alpha, i = 0, 1, \dots, n\}. \quad (10)$$

In this case, we claim that the data stream changes abnormally after time point τ , and the process enters into an out-of-control state.

2.4. Fault Diagnosis Procedure

To this end, we first keep monitoring a p -dimensional data stream by using the monitoring statistics $\tilde{M}_t$. After the OC alarm is issued, we enter into the second stage of statistical fault diagnosis to identify the components that have undergone abnormal changes. To this end, we continue to monitor the data stream for a while, and collect some more OC observations. The diagnostic problem can be regarded as the following multiple hypothesis testing. The null hypothesis and the alternative hypothesis are:

$$\mathcal{H}_{0jt} : \mu_{1j} - \mu_{0j} = 0, \text{ v.s. } \mathcal{H}_{1jt} : \mu_{1j} - \mu_{0j} \neq 0, \text{ for } j = 1, \dots, p, t > \hat{\tau}. \quad (11)$$

It is obviously that if $\mathcal{H}_{0jt}$ is accepted, the j -th component of the stream is IC, otherwise, a rejection of $\mathcal{H}_{0jt}$ indicates that the j -th component of the stream is OC when $t > \hat{\tau}$.

To address this problem, a diagnostic statistic $\tilde{W}_{jt}$ is proposed as the following form,

$$\tilde{W}_{jt} = \frac{\bar{y}_{jt}^2}{\frac{\gamma}{n(2-\gamma)}\hat{\sigma}_{jj}}, \quad j = 1, \dots, p, t > \hat{\tau}, \quad (12)$$

where $\bar{y}_{jt} = \frac{1}{n} \sum_{i=\hat{\tau}+1}^{\hat{\tau}+n} y_{it}$ is the mean value of the observed samples, the sample variance $\hat{\sigma}_{jj}$ is defined as that in Equation (6). It can be seen that if the value of $\tilde{W}_{jt}$ is large, it indicates that abnormal changes may occur in the data stream. Therefore, it can be inferred that the null hypothesis in Equation (11) is rejected if $\tilde{W}_{jt}$ too large to exceed a threshold.

In many applied multiple testing problems, there are many test strategies, such as the traditional critical value test methods and some FDR controlling procedures [19?, 20]. It is obvious that $E(\bar{y}_{jt}) = 0$, $Var(\bar{y}_{jt}) = \frac{\gamma}{n(2-\gamma)}\hat{\sigma}_{jj}$, and $\frac{\bar{y}_{jt}}{\sqrt{\frac{\gamma}{n(2-\gamma)}\hat{\sigma}_{jj}}} \sim N(0, 1)$ as $n \rightarrow \infty$ under the null hypothesis $\mathcal{H}_{0jt}$. Thus, the diagnostic statistic $\tilde{W}_{jt}$ follows Chi-squared distribution,

$$\tilde{W}_{jt} \rightarrow \chi^2(1), \text{ as } n \rightarrow \infty. \quad (13)$$

In the diagnosis process, we should ensure that the global per-comparison error rate (PCER) can be controlled within a prespecified level α . Without loss of generality, we assume the first p_0 components of the high-dimensional observations are IC, and the remaining $p - p_0$ components are

OC when $t = \tau + 1, \tau + 2, \dots, \tau + n$. The PCER can be defined as the expected number of Type-I errors divided by the number of hypotheses, that is,

$$\text{PCER}_t = \frac{1}{p} \sum_{j=1}^{p_0} P_{H_{1t}}(\tilde{W}_{jt} > c_h | t > \hat{\tau}) \leq \alpha. \quad (14)$$

Therefore, we need to determine c_h such that the Equation (14) is satisfied, the choice of c_h can be determined through numerical simulation. More specifically, we let the global per-comparison error rate equals α , that is,

$$\frac{1}{p} \sum_{j=1}^p P_{H_{1t}}(\tilde{W}_{jt} > c_h | t > \hat{\tau}) = \alpha. \quad (15)$$

It is obvious that when Equation (15) is satisfied, the Equation (14) above is naturally controlled within α and holds true.

For a given initial threshold value of c_h , (such as $c_h = \chi^2_{1-\alpha}(1)$), the proportion of $\tilde{W}_{jt} > c_h$ is denoted as $\hat{p}$ when $t > \hat{\tau}$. Then, we repeat it B times ($B = 2000$), the average values of $\hat{p}$ over B replications can be used to approximate $\frac{1}{p} \sum_{j=1}^p P_{H_{1t}}(\tilde{W}_{jt} > c_h | t > \hat{\tau}) = \alpha$. Conversely, if Equation (15) is satisfied, we use numerical methods to obtain an approximate value, denoted as $\hat{c}_h$. Therefore, if $\tilde{W}_{jt}$ is not larger than $\hat{c}_h$, we can control the the global per-comparison error rate at a level α , otherwise if $\tilde{W}_{jt}$ is larger than $\hat{c}_h$, this indicates that the j -th component of the data stream is abnormal.

2.5. Algorithm Steps

The main algorithm are summarized as the following steps.

Algorithm 1

Online Monitoring Stage	
1	Given in-control data streams X_{ic} ;
2	Calculate the sample mean $\bar{X}_{ic}$ and standard deviation $std(X_{ic})$ of the IC observations;
3	Observe p -dimensional data stream X_t , the EWMA statistic as calculate $\tilde{Y}_t = \gamma(X_t - \bar{X}_{ic}) + (1 - \gamma)Y_{t-1}$, $t = 1, 2, \dots$;
4	Construct global monitoring statistic $\tilde{M}_t = \max_{1 \leq j \leq m} \frac{\tilde{y}_{jt}^2}{\frac{\gamma}{2-\gamma} \hat{\sigma}_{jj}}$;
5	If $\tilde{M}_t > L_\alpha$, then raise the OC alarm, the estimated change point is $\hat{\tau} = \inf\{t : t > t_{ic} \tilde{M}_{t+i} > L_\alpha, i = 0, 1, \dots, n\}$.
Fault Diagnosis Stage	
1	Collect n observations after the OC alarm signal occurs;
2	Construct diagnostic statistic $\tilde{W}_{jt} = \frac{\tilde{y}_{jt}^2}{\frac{\gamma}{n(2-\gamma)} \hat{\sigma}_{jj}}$, $j = 1, \dots, p$, $t = 1, 2, \dots, \hat{\tau} + n$;
3	Record the observations set with $\tilde{M}_t \leq L_\alpha$ for $t = 1, 2, \dots, \hat{\tau} + n$ as H , let $\hat{p} = \frac{1}{p} \sum_{j=1}^p P_{H_{0t}}(\tilde{W}_{jt} > c_h \tilde{M}_t > L_\alpha)$;
4	Randomly sample B times from data set H , the average of $\hat{p}$ over B replications is used to approximate $\hat{c}_h$;
5	Obtain $\hat{\mathcal{J}}_t = \{j : \tilde{W}_{jt} > \hat{c}_h \tilde{M}_t > L_\alpha\}$, the diagnosed variables are the ones corresponding to the indicators of $\hat{\mathcal{J}}_t$.

2.6. Performance Evaluation Measures

In the online monitoring phase, in order to measure the performance of the proposed monitoring method, we choose the following three metrics. The first is the accuracy of the change point estimation, that is, when the estimated value $\hat{\tau}$ is close to the real change point τ , the change point estimation is accurate. The second is the Type-I error rate of online monitoring, which is the proportion of normal

observations mistakenly identified as outliers. The last one is the power value of online monitoring, that is, the proportion of abnormal observations that are accurately identified. The Type-I error rate can be well controlled, meanwhile, the larger of the power value, approaching 100%, the better of the proposed monitoring method.

In the fault diagnosis stage, we define two relevant metrics to verify the effectiveness of the method, one is the false positive rate (FPR), a measure of the proportion of not-faulty variables that are incorrectly identified as faulty variables. The other is the true positive rate (TPR), defined as the percentage of the ratio of number of variables that are correctly detected as faulty over the number of all not-faulty variables. Without loss of generality, after an change point τ , the IC components of the stream are denoted as $\mathcal{J}^{IC} = \{j : j = 1, 2, \dots, p_0\}$, and the OC components set is $\mathcal{J}^{OC} = \{j : j = p_0 + 1, p_0 + 2, \dots, p\}$. At any time point t , the estimated OC components are defined as $\hat{\mathcal{J}}_t = \{j : \hat{W}_{jt} > \hat{c}_h | \hat{M}_t > L_\alpha\}$. Thus, the TPR and FPR values at time point t are defined as

$$TPR_t = \frac{|\hat{\mathcal{J}}_t \cap \mathcal{J}^{OC}|}{|\mathcal{J}^{OC}|} \cdot 100\%; \quad FPR_t = \frac{\max\{0, |\hat{\mathcal{J}}_t - \mathcal{J}^{OC}|\}}{|\mathcal{J}^{IC}|} \cdot 100\%.$$

where $|A|$ represents the number of elements in set A . Correspondingly, up to time point t , the average TPR and FPR values can be defined as follows,

$$TPR = \frac{1}{t - \hat{\tau}} \sum_{i=\hat{\tau}+1}^t Power_i; \quad FPR = \frac{1}{t - \hat{\tau}} \sum_{i=\hat{\tau}+1}^t Size_i.$$

In fact, we should ensure that the higher the average TPR value, the better; meanwhile, we should also ensure that the FPR value, the smaller, the better the performance.

3. Simulation Studies

In this section, we evaluate the proposed monitoring and diagnosis procedure through a simulation study, and compare it with some existing methods. In the experiments, the IC data stream follows $N(\mu_0, \Sigma_p)$, while the out-of-control stream follows $N(\kappa\mu_1, \Sigma_p)$, where κ is a parameter used to measure the degree of drift. For convenience, we assume that μ_0 is equal to the $\mathbf{0}$ vector here, μ_1 is a sparse p -dimensional mean vector, with most components equal to zero and a few components equal to 1. That is, we assume that the abnormal change occurs after a time point τ , when $t < \tau$, the process is IC, while $t > \tau$, the p_s components of its mean vector are equal to κ , and the remaining $p - p_s$ components are all equal to 0.

The covariance matrix Σ_p , which characterizes the correlation between p -dimensional variables. In the simulation study, we assume that Σ_p remains unchanged before and after the change occurs, the following three covariance matrix structures are considered.

- (a) Independent case, $\Sigma_p = I_{p \times p}$ is an identity matrix;
- (b) Long-range correlation case, $\Sigma_p = (\rho^{|k-l|})_{p \times p}$, $k, l = 1, \dots, p$, and $\rho = 0.5$ is chosen;
- (c) Short-range correlation case, Σ_p is a block-diagonal matrix, within each block matrix, $\Sigma_d = (\rho^{|k-l|})_{d \times d}$, $k, l = 1, \dots, d$, where d is the block size and $\rho = 0.6$, $d = 10$.

The method proposed in this paper is based on extreme value theory (EVT), so we refer to it as the EVT method. In addition, we simulated two other methods and compared their monitoring and diagnostic performance. The first method uses an adaptive principal component (APC) selection approach based on hard-thresholding to select the most susceptible sets according to an unknown change [5]; The second method proposed by Amir and Mohsen [4], which contains two steps, a single Hotelling chi-square chart and a Shewhart chart, we refer to it as the AM method. Thus, in the simulation study, we use these methods to monitor data streams, diagnose anomalous variables, and compare simulation results. All the results are obtained from 1,000 replicated simulations.

Table 1 shows the simulated type-I error and power values in the online monitoring phase by the proposed method, the standard deviations are included in the corresponding brackets. As shown in

the table, when $t > \tau$, the proportion p_s/p ranges from 10% to 25%, the empirical type-I error values could be controlled around 5% in most cases. Moreover, with the increase of κ , the power value also increases. This phenomenon can be easily explained by the fact that, as the degree of abnormal data increases, making abnormal data more easily monitored.

Table 1. Average percentage of the empirical type-I error values $\tilde{\alpha}$ (%) and the power values $\tilde{\beta}$ (%) by the proposed monitoring procedure under different model settings, when $p = 200, n_{ic} = 150, t = 500, \tau = 200, \alpha = 0.05$ and $\kappa = 1.5, 2.0, 2.5$ respectively.

Models	p_s/p	$\kappa=1.5$		$\kappa=2.0$		$\kappa=2.5$	
		$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$
(a)	0.05	5.3(3.4)	99.2(0.5)	5.4(3.4)	99.7(0.2)	5.5(3.5)	99.9(0.2)
	0.10	5.2(3.4)	99.7(0.2)	5.4(3.4)	99.8(0.7)	5.5(3.4)	99.9(0.1)
	0.20	5.6(3.6)	99.8(0.2)	5.6(3.1)	99.9(0.1)	5.7(3.6)	100.0(0.1)
(b)	0.05	5.5(3.2)	96.6(1.4)	5.4(3.6)	99.7(0.2)	5.2(3.2)	99.9(0.2)
	0.10	5.4(3.4)	99.5(0.3)	5.3(3.7)	99.8(0.2)	4.9(3.4)	99.9(0.1)
	0.20	5.4(3.6)	99.7(0.2)	5.3(3.5)	99.8(0.2)	5.3(3.3)	100.0(0.1)
(c)	0.05	4.9(3.1)	96.7(1.3)	5.3(3.3)	99.7(0.2)	5.3(3.2)	99.9(0.2)
	0.10	5.1(3.5)	99.6(0.3)	5.3(3.6)	99.8(0.2)	5.4(3.5)	99.9(0.1)
	0.20	5.1(3.2)	99.7(0.2)	5.4(3.5)	99.9(0.2)	5.4(3.4)	100.0(0.1)

Figure 2 shows the change trend of the proposed monitoring statistics over time points in different model scenarios. In the three subfigures on the left, the proportion of abnormal data flow is 5%, while in the three subplots on the right, the proportion of abnormal data flow is 20%. In the high-dimensional data stream being monitored, the real change point time $\tau = 200$. From the results of Figure 2, it can be seen that in all three model scenarios, the values of monitoring statistics show a significant upward trend after the change point occurs. Compared to the three subplots on the left, the value of the monitoring statistic on the right shows a more significant upward trend after the change point occurs. This is because the graph on the right depicts the situation where abnormal data accounts for 20%, and the value of this monitoring statistic will increase significantly. It is worth noting that before the change point occurs, the values of monitoring statistics may exceed the threshold at certain time points, which is caused by outliers and is not surprising. In summary, from Figure 2, it can be seen that this monitoring statistic can effectively reflect the changing trend of high-dimensional data streams.

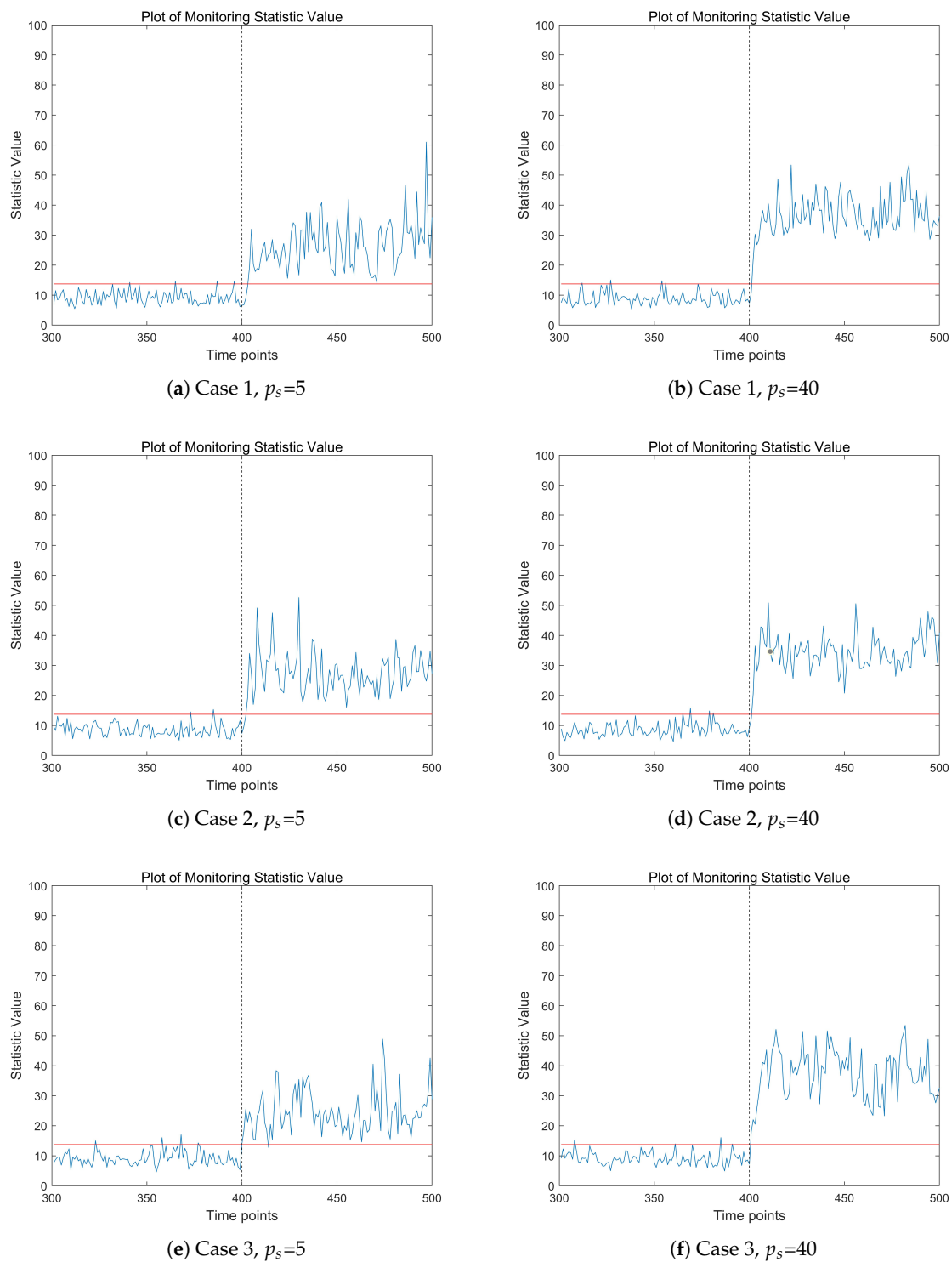


Figure 2. The change trend of the monitoring statistic value, when $p = 200$, $\tau = 400$, $t = 500$, $\kappa = 2.5$. The red dashed line represents the threshold line.

In Table 2, we present the monitoring results of three methods and compare them by estimating the time points of change, standard deviation, and the difference between the true and estimated values. From the results, it can be seen that, at a monitoring duration of $t = 500$, when the true change point τ is 200 or 400, as κ increases from 1.0 to 2.5, the difference and standard deviation between the estimated value and the true value gradually decrease, resulting in a large number of repeated simulations, and the probability of the difference being less than 3 is close to 100%.

Table 2. Comparison of the three methods via change point estimation with $p = 200, t = 500, \alpha = 0.05$, the real change point $\tau = 200$ and 400 respectively. Bias: the absolute deviation of estimators; Sd: the standard deviations of estimators; $P_3 = 100\Pr(|\tau - \hat{\tau}| \leq 3)\%$.

Model	Method	κ	$\tau=200$			$\tau=400$		
			Bias	Sd	P_3	Bias	Sd	P_3
(II)	EVT	1.0	6.3	3.8	24.0	7.1	5.0	23.6
		1.5	2.1	1.2	89.0	2.2	1.2	88.5
		2.0	1.1	1.1	99.5	1.2	0.8	99.5
		2.5	0.7	0.7	100.0	0.6	0.7	100.0
	A-M	1.0	9.4	4.5	13.5	10.3	3.7	5.6
		1.5	6.1	2.8	17.5	6.8	2.5	7.6
		2.0	4.1	2.1	34.0	4.8	2.0	17.6
		2.5	3.4	1.5	47.5	3.9	1.5	33.6
	APC	1.0	11.4	8.6	15.1	6.9	6.7	33.4
		1.5	5.2	3.2	33.5	3.1	1.7	66.5
		2.0	3.6	1.9	56.5	2.3	0.9	92.6
		2.5	2.9	1.1	78.4	1.8	0.7	97.5
(III)	EVT	1.0	6.9	4.8	27.5	8.0	5.1	15.5
		1.5	2.4	1.3	81.5	2.3	1.4	86.0
		2.0	1.3	0.8	100.0	1.0	0.9	97.5
		2.5	0.9	0.7	100.0	0.8	0.7	100.0
	A-M	1.0	12.0	3.9	4.0	13.7	2.9	3.0
		1.5	7.8	2.7	6.5	8.6	2.3	4.5
		2.0	5.8	1.9	10.5	6.5	1.9	10.0
		2.5	4.7	1.4	17.0	5.0	1.4	16.5
	APC	1.0	15.4	12.3	13.5	8.7	9.9	33.5
		1.5	6.9	4.9	25.5	4.5	3.3	49.5
		2.0	4.4	2.7	51.5	2.5	1.5	78.5
		2.5	3.0	1.3	76.5	1.9	0.9	91.5

As is well known, the size of γ can have an impact on the monitoring results of data streams, especially the EWMA statistic. So, in Table 3, we simulated the results of the first type of error, monitoring effectiveness, and change point estimation under different gamma value selections. From the simulation results in Table 3, it can be seen that when the gamma values are 0.2, 0.4, and 0.6, the type-I error values remain around 5%. However, as γ increases, the monitoring power values decrease and the change point estimation value also deteriorates. Meanwhile, as κ increases, the different choice of γ have little effect on the results, which confirms that the EWMA statistic is sensitive to small mean drifts.

Table 3. Average percentage of type-I error values ($\tilde{\alpha}\%$), power values ($\tilde{\beta}\%$) and change point estimators ($\hat{\tau}$) by the proposed online monitoring procedure under different choices of γ , when $p = 200$, $t = 200$, $\tau = 100$ and $\alpha = 0.05$.

p_s/p	γ	$\tilde{\alpha}$	$\kappa=1.0$		$\hat{\tau}$	$\kappa=1.5$		$\hat{\tau}$	$\kappa=2.0$		$\hat{\tau}$
			$\tilde{\beta}$	$\hat{\tau}$		$\tilde{\beta}$	$\hat{\tau}$		$\tilde{\beta}$	$\hat{\tau}$	
0.05	0.2	4.8	80.4	209.5	4.9	97.1	202.8	4.9	98.4	201.4	
	0.4	4.9	34.7	265.1	5.0	83.4	207.3	4.9	98.3	201.5	
	0.6	5.0	17.8	290.1	5.0	50.3	248.1	4.9	85.9	205.2	
0.10	0.2	5.0	92.9	206.0	5.0	97.9	201.6	4.9	98.8	200.8	
	0.4	5.0	53.9	232.9	4.9	95.7	202.5	5.0	99.2	200.8	
	0.6	4.9	28.1	292.5	5.0	73.4	213.1	4.9	97.4	201.5	
0.15	0.2	5.0	95.9	204.1	5.0	98.3	201.5	4.9	99.0	200.4	
	0.4	4.9	67.1	215.6	5.0	98.2	201.6	4.9	99.2	200.6	
	0.6	5.0	37.3	281.7	5.0	84.7	205.5	4.9	99.2	200.5	

Figure 3 simulates the trend of asymptotic power values and the Type-I error values obtained by three different monitoring methods with the increase of κ under the three model settings. From Figure 3, it can be seen that on the left side, when $\kappa = 0.4$ or $\kappa = 0.8$, the power value obtained by the AM method is slightly better than our method, but, when κ increases to 1.2 or even greater, the power values obtained by our EVT method are significantly better than the other two methods. On the right of Figure 3, it can be seen that the type-I error obtained using the EVT method can be effectively controlled at around 5%, while the AM method cannot control the type-I error effectively in model (I). The results by the APC method also have poor performance,

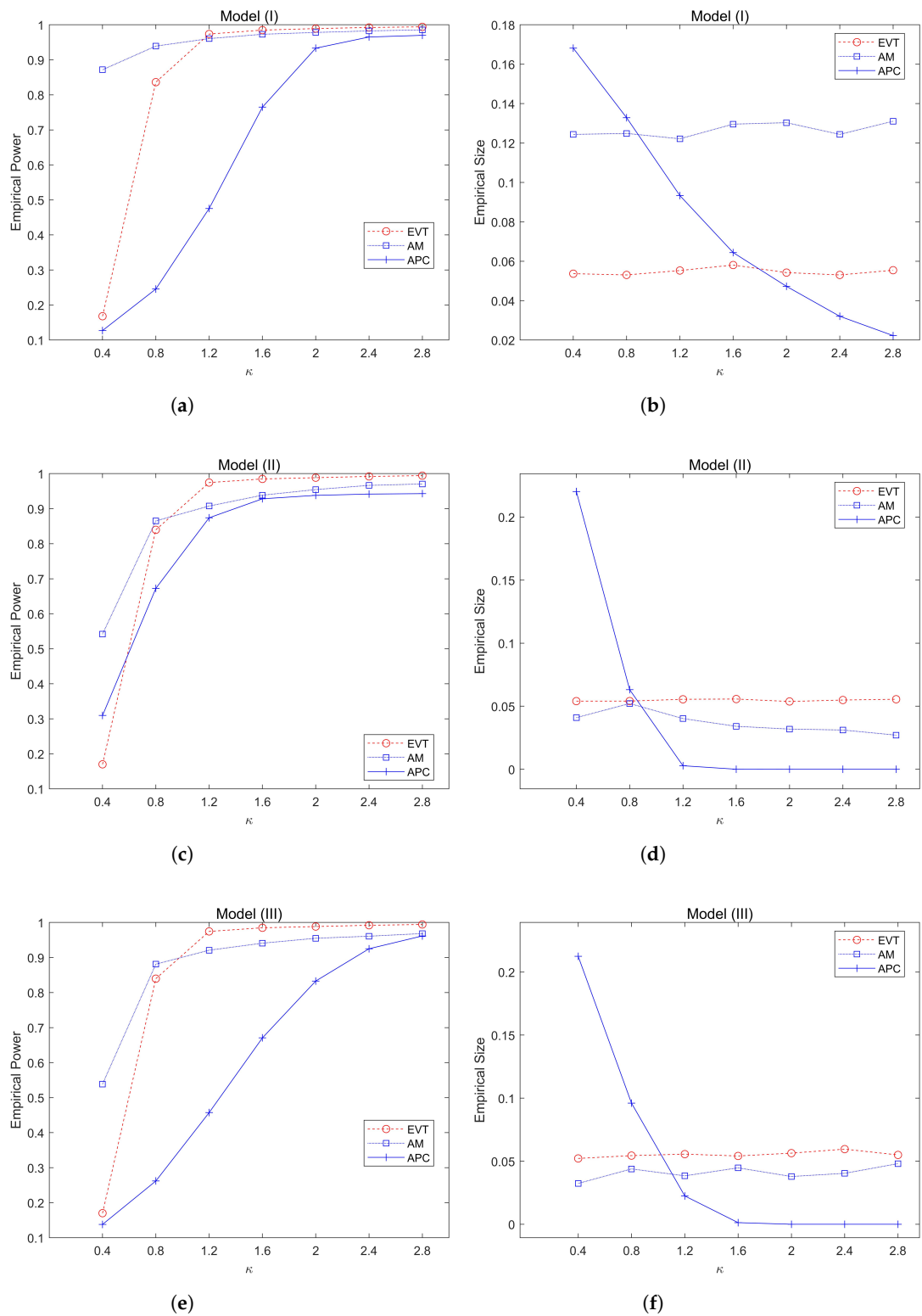


Figure 3. The change trend of power and size values (%) by various procedures when $p = 200$, $p_s = 20$, $\tau = 200$, $t = 300$, $\alpha = 0.05$ with the increase of the parameter κ .

Table 4 lists the fault diagnosis results of high-dimensional data streams by using different diagnostic methods. We have calculated the values of two main indicators here, one is the correct recognition rate, and the other is the incorrect recognition rate. At the same time, we also considered the value of change point estimation during the online monitoring phase. The results show that as k increases, the change point estimation becomes closer to the true value. Regardless of the proportion of

p_s being equal to 5%, 10%, or 15%, the EVT method is significantly better than the other two methods. Furthermore, from the diagnostic results, the TP value of the EVT method is close to 100%. It should be noted that the AM method and APC method also have good diagnostic results, but considering the results of point estimation, overall, our method is superior to the other two methods.

Table 4. Diagnosis accuracy of the proposed procedure, under different choices of κ and p_s , when $m = 10$, $p = 100$, $t = 200$, $\tau = 100$ and $\alpha = 0.05$.

		EVT			A-M			APC		
p_s	κ	$\hat{\tau}$	TPR	FPR	$\hat{\tau}$	TPR	FPR	$\hat{\tau}$	TPR	FPR
5%	1.0	212.3	99.1	0.8	213.6	100.0	0.9	210.5	71.1	4.4
	1.5	203.3	100.0	0.8	208.6	100.0	0.9	208.1	82.1	5.4
	2.5	201.6	100.0	0.8	206.7	100.0	0.9	206.9	85.8	7.4
10%	1.0	206.5	99.3	1.3	209.5	100.0	0.9	209.9	84.8	10.3
	1.5	202.1	100.0	1.4	206.1	100.0	1.0	206.7	92.8	12.9
	2.5	201.2	100.0	1.3	204.6	100.0	0.8	206.2	94.4	15.0
15%	1.0	204.5	99.5	1.6	209.8	100.0	0.6	208.1	91.8	12.6
	1.5	201.4	99.7	1.7	205.0	100.0	0.9	206.2	95.7	14.2
	2.5	200.9	100.0	1.7	204.4	100.0	0.4	205.5	96.1	13.4

4. Real Data Analysis

In order to verify the effectiveness of the method proposed in this article, we apply the proposed monitoring and diagnostic methods to analyze the real data set. The glass data contains Electron Probe X-ray Microanalysis (EPXMA) intensities for different wavelengths of sixteenth to seventeenth century archaeological glass vessels [21]. We consider this data as a high-dimensional data stream sorted by time points, with 180 time points and each observation containing 750 variables at each time point. Many literatures in abnormal point detection have analyzed this data set, such as Hubert et al. (2005)[22] performed the RobPCA method to analysis the glass data, and identified some row-wised outliers. Hubert et al. (2019)[23] also analyzed this data set by the proposed MacroPCA procedure, and flagged some outlying samples as outliers. In this section, we perform the proposed monitoring and diagnosis procedures on the glass data. Firstly, through preliminary data analysis, we found that 13 variables were missing or invalid. Therefore, for each observation, we first removed 13 variables and left 737 variables for further analysis. Thus, we obtained 180 high-dimensional time series data streams.

In the first stage, our goal is to monitor the time series data stream online and accurately detect abnormal data. So, we selected appropriate parameters and applied the online monitoring strategy proposed in this paper to the data stream, obtaining the results shown in the following figure. The results showed that from time point 57 onwards, there was a loss of control in the data flow, which is consistent with Hubert’s research conclusion.

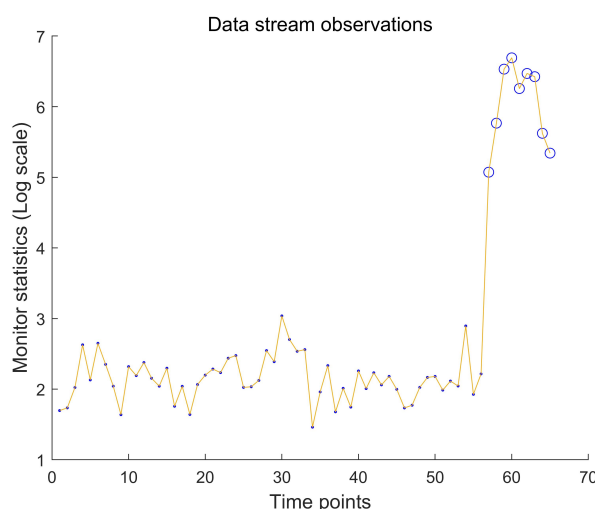


Figure 4. The data stream with IC observations (Solid Points), and OC observations (Hollow Circle) when monitoring data streams.

In the second stage, we apply the statistical diagnostic method proposed in this article to further analyze the out of control data stream and identify the variables that cause the data flow to become out of control. During the online monitoring phase, it can be observed that when $t > 57$, the data stream experienced a loss of control. Therefore, we allow the data stream to run for a certain time period (the time length is denoted as 5), and then performed fault diagnosis on the observed values that experienced the loss of control. According to the diagnostic analysis results, when the observation time point of the data stream is greater than 57, out of a total of 737 variables, 36 variables were detected to have undergone abnormal changes, resulting in abnormal drift of the data stream.

5. Conclusions

In this article, we studied the online monitoring and fault diagnosis of high-dimensional data streams. In the first stage, we proposed a online monitoring procedure based on EWMA statistics. As is well known, the EWMA statistic is more sensitive to small changes and it includes information from previous samples. Thus, an EWMA statistic is chosen in this study for monitor high-dimensional stream. We have constructed online monitoring statistics and provided corresponding online monitoring solutions. In the second stage, we constructed diagnostic statistics and provided corresponding statistical diagnostic algorithms. Finally, we validated the effectiveness of the method through numerical simulations and case studies. The simulation results show that our method can quickly and effectively monitor the timing of change points, and can effectively control the rate of erroneous monitoring. Compared with other existing methods, the proposed online monitoring and diagnosis methods have great advantages in controlling type-I errors and also has good power for identifying outlying points, especially for high-dimensional sparse data stream.

References

1. Zou, C.; Wang, Z.; Zi, X.; Jiang, W. An efficient online monitoring method for high-dimensional data streams. *Technometrics* **2015**, *57*, 374–387.
2. Li, J. A two-stage online monitoring procedure for high-dimensional data streams. *J. Qual. Technol.* **2019**, *51*, 392–406.
3. Li, W.; Xiang, D.; Tsung, F.; Pu, X. A diagnostic procedure for high-dimensional data streams via missed discovery rate control. *Technometrics* **2020**, *62*, 84–100.
4. Ahmadi-Javid, A.; Ebadi, M. A two-step method for monitoring normally distributed multi-stream processes in high dimensions. *Qual. Eng.* **2020**, *33*, 143–155.
5. Ebrahimi, S.; Ranjan, C.; Paynabar, K. Monitoring and root-cause diagnostics of high-dimensional data streams. *J. Qual. Technol.* **2022**, *54*, 20–43.

6. Liu, K.; Mei, Y.; Shi, J. An Adaptive Sampling Strategy for Online High-Dimensional Process Monitoring. *Technometrics* **2015**, *57*, 305–319.
7. Huang, T.; Kandasamy, N.; Sethu, H.; Stamm, M. An Efficient Strategy for Online Performance Monitoring of Datacenters via Adaptive Sampling. *IEEE Trans. Cloud Comput.* **2019**, *7*, 155–169.
8. Xiang, D.; Li, W.; Tsung, F.; Pu, X.; Kang, Y. Fault classification for high-dimensional data streams: A directional diagnostic framework based on multiple hypothesis testing. *Nav. Res. Logist. (NRL)* **2021**, *68*, 973–987.
9. Du, L.; Zou, C. On-line control of false discovery rates for multiple datastreams. *J. Stat. Plan. Inference* **2018**, *194*, 1–14.
10. Li, J. Efficient global monitoring statistics for high-dimensional data. *Qual. Reliab. Eng. Int.* **2020**, *36*, 18–32.
11. Xian, X.; Zhang, C.; Bonk, S.; Liu, K. Online monitoring of big data streams: A rank-based sampling algorithm by data augmentation. *J. Qual. Technol.* **2021**, *53*, 135–153.
12. Xiang, D.; Qiu, P.; Wang, D.; Li, W. Reliable Post-Signal Fault Diagnosis for Correlated High-Dimensional Data Streams. *Technometrics* **2022**, p. 64.
13. Colosimo, B.; Jones-Farmer, L.; Megahed, F.; Paynabar, K.; Ranjan, C.; Woodall, W. Statistical Process Monitoring from Industry 2.0 to Industry 4.0: Insights into Research and Practice. *Technometrics* **2024**, *66*, 1–35. <https://doi.org/10.1080/00401706.2024.2327341>.
14. Hu, X.; Zhao, Y.; Yeganeh, A.; Shongwe, S. Two memory-based monitoring schemes for the ratio of two normal variables in short production runs. *Comput. Ind. Eng.* **2024**, *198*, 110690. <https://doi.org/10.1016/j.cie.2024.110690>.
15. Yoav, B.; Daniel, Y. A multivariate exponentially weighted moving average control chart. *Ann. Stat.* **2001**, *29*, 1165–1188.
16. 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.
17. Feng, L.; Ren, H.; Zou, C. A setwise EWMA scheme for monitoring high-dimensional datastreams. *Random Matrices: Theory Appl.* **2020**, *9*, 2050004.
18. Cai, T.T.; Liu, W.; Xia, Y. Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc.* **2014**, *76*, 349–372.
19. Yoav, B.; Yosef, H. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)* **1995**, *57*, 289–300.
20. Du, L.; Guo, X.; Sun, W.; Zou, C. False Discovery Rate Control Under General Dependence By Symmetrized Data Aggregation. *J. Am. Stat. Assoc.* **2021**, *118*, 1–34. <https://doi.org/10.1080/01621459.2021.1945459>.
21. Lemberge, P.; De Raedt, I.; Janssens, K.H.; Wei, F.; Van Espen, P.J. Quantitative analysis of 16–17th century archaeological glass vessels using PLS regression of EPXMA and μ -XRF data. *J. Chemom. A J. Chemom. Soc.* **2000**, *14*, 751–763.
22. Hubert, M.; Rousseeuw, P.; Branden, K. ROBPCA: A new approach to robust principal component analysis. *Technometrics* **2005**, *47*, 64–79. <https://doi.org/10.1198/004017004000000563>.
23. Hubert, M.; Rousseeuw, P.; Van den Bossche, W. MacroPCA: An All-in-One PCA Method Allowing for Missing Values as Well as Cellwise and Rowwise Outliers. *Technometrics* **2019**, *61*, 459–473. <https://doi.org/10.1080/00401706.2018.1562989>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.