# Reliable Post-Signal Fault Diagnosis for Correlated High-Dimensional Data Streams

Dongdong Xiang[1], Peihua Qiu[2], Dezhi Wang[1,3], Wendong Li[4]

[1] KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China

[2]Department of Biostatistics, University of Florida, Gainesville, USA

[3]School of Mathematics and Statistics, Lanzhou University, Lanzhou, China

[4]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

**Abstract**

Rapid advance of sensor technology is facilitating the collection of high-dimensional data streams (HDS). Apart from real-time detection of potential out-of-control (OC) patterns, post-signal fault diagnosis of HDS is becoming increasingly important in the filed of statistical process control to isolate abnormal data streams. The major limitations of the existing methods on that topic include i) they cannot achieve reliable diagnostic results in the sense that their performance is highly variable, and ii) the informative correlation among different streams is often neglected by them. This paper elaborates the problem of reliable fault diagnosis for monitoring correlated HDS using the large-scale multiple testing. Under the framework of hidden Markov model dependence, new diagnostic procedures are proposed, which can control the missed discovery exceedance (MDX) at a desired level. Extensive numerical studies along with some theoretical results show that the proposed procedures can control MDX properly, leading to diagnostics with high reliability and efficiency. Also, their diagnostic performance can be improved significantly by exploiting the dependence among different data streams, which is especially appealing in practice for identifying clustered OC streams.

**Keywords**: Fault diagnosis; Hidden Markov models; Missed discovery exceedance; Multiple tests; Reliability; Statistical process control.

Corresponding author: Wendong Li, wendongli01@gmail.com

# 1 Introduction

With the rapid development of modern sensor and data acquisition technologies, high-dimensional data streams (HDS) that involve large-scale continuous sequential streaming data have become ubiquitous nowadays, which brings tremendous challenges to the field of multivariate statistical process control (SPC). In such a situation, if a complex high-dimensional process operates abnormally, engineers are paying more attention to the identification of the data streams that are responsible for the anomaly condition in order to locate and eliminate the root cause of the problem. This task is typically referred to as fault diagnosis. While online monitoring of HDS has attracted a considerable attention recently (Liu et al., 2015; Zou et al., 2015; Yan et al., 2018; Zhang et al., 2020; Li et al., 2020b), high-dimensional fault diagnosis is still an active area with great potential.

In the literature of conventional multivariate SPC, post-signal fault diagnosis often relies heavily on the assumption that the dimensionality of the underlying process is low-to-moderate (Qiu 2014). Early works in this area tried to capture the relationship among variables by interpreting and decomposing Hotelling's $T^2$-type statistics (Mason, et al., 1995, 1997; Li et al., 2008), based on which many step-down procedures were further proposed (Sullivan et al., 2007; Zhu and Jiang, 2009; Kim et al., 2016). Recently, one trend is to reduce the dimension into a smaller number by using variable selection techniques. Zou et al. (2011) introduced a multivariate diagnostic framework that integrates the Bayesian information criterion (BIC) with the adaptive LASSO algorithm, which has been shown to have a better performance than many conventional methods. Many variable-selection-based control charts have also been proposed (Wang and Jiang, 2009; Zou and Qiu, 2009; Capizzi and Masarotto, 2011; Li et al., 2017). Although they are designed mainly for online monitoring, they could be used for a rough fault diagnosis by regarding the selected variables as out-of-control (OC) streams.

To handle the fault diagnosis problem for HDS, the conventional multivariate approaches may fail. Specifically, when the dimensionality of the process is extremely high, the "curse of dimensionality" arises, leading to a poor diagnostic performance. Besides, these approaches are computationally too intensive to afford when they are applied to HDS. See Zou et al. (2011) and Li et al. (2020a) for detailed discussions. More recently, in order to make fault

diagnosis methods practically useful, Zhang et al. (2020) proposed a diagnostic framework based on the square-root LASSO algorithm. Another seminal work is Li et al. (2020a), who handled the problem of fault diagnosis for HDS based on large-scale multiple testing, and controlled the weighted missed discovery rate (MDR) at a desired level while minimizing the expected number of false positives. This multiple testing framework exhibits a higher diagnostic power and a better computational efficiency, compared to the conventional methods. Xiang et al. (2021) further proposed a directional diagnostic approach based on a three-classification multiple-testing framework to determine the shift directions systematically.

Unfortunately, most methods mentioned above share two major drawbacks. First, they do not take into account the between-stream correlation properly, and some even assume that the data streams are independent of each other, which is rarely valid in practice. For HDS, between-stream correlation often exists, and the data streams physically located close to the OC ones would be more likely to be OC. If such information is not considered systematically, the resulting methods would suffer from a substantial loss of information and sacrifice their diagnostic power. Second, these existing methods try to achieve their diagnostic goals from the perspective of a large-sample theory in the sense that their diagnostic accuracy can only be controlled on the basis of a large number of replicated diagnoses. For any single diagnosis, the result could be highly variable. For example, in Li et al. (2020a), the target accuracy measure to control is MDR, which is defined to be the mathematical expectation of the missed discovery proportion (MDP, i.e., the proportion of false negatives among all the OC streams in a single diagnosis). When MDR equals $\alpha$, it means that the mathematical expectation of MDP equals $\alpha$, but the probability of MDP being larger than $\alpha$ in a single diagnosis can still be very high, which will be shown in details later. In such a situation, these methods may have a poor diagnostic performance in a single diagnosis.

In conclusion, so far there are no diagnostic procedures for HDS with a high reliability that can accommodate the between-stream correlation. To fill the gap, this paper proposes a reliable fault diagnostic procedure based on a novel accuracy measure, called missed discovery exceedance (MDX), which is defined to be the probability that MDP exceeds a given value $\tau$. By controlling MDX at a desired level $\alpha$, the probability that MDP is less than $\tau$ would be $1-\alpha$, which ensures a high diagnostic reliability. To properly model the correlation structure among different streams, we suggest using a hidden Markov model (HMM), which has been

shown an effective tool and widely used in many areas, including signal processing, speech recognition, DNA sequence analysis, and influenza-like illness analysis (Ephraim and Merhav, 2002). In all these problems, the related variables are ordered by their spatial or procedural locations, and their correlation is often described by a HMM model. Then, the fault diagnosis problem for HDS under the HMM dependence structure is investigated in a novel MDX-based multiple testing framework. To this end, an oracle diagnostic procedure is first proposed under the assumption that all process parameters are known. Then, by plugging in consistent estimates of the unknown parameters, a data-driven diagnostic procedure is proposed which mimics the oracle one. Our method is different from the conventional diagnostic methods in that it takes into account the large variability of MDP in a single diagnosis by controlling the level of MDX, and is highly desirable in cases when the data streams are mutually correlated by the use of HMM. Extensive numerical studies will show that it has a satisfactory diagnostic performance in various cases considered.

To illustrate the major feature of the proposed method, let us consider the following toy example. Let $(\theta_i)_1^m \in \{0,1\}^m$ be a Markov chain with the initial state 0 and the transition matrix

$$\begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix},$$

and $\boldsymbol{x} = (x_1, ..., x_m)^T$ be an $m$-dimensional OC observation following the mixture model $x_i|\theta_i \sim (1-\theta_i)N(0,1) + \theta_i N(2,1)$, for each $i$, where $m = 1000$. A comparison of the MDR-based procedure by Li et al. (2020a) and the proposed oracle MDX-based procedure described in Section 3 is shown in Figure 1, where the density functions of the MDP levels of the two methods are shown. In this example, both $\alpha$ and $\tau$ are set at 0.05 for simplicity. The vertical solid line denotes the $100(1-\alpha)$th percentile of MDP of the MDX-based procedure, and the vertical dashed line is the 50th percentile of MDP of the MDR-based procedure. It can be seen that the chance for MDP to be above 0.05 is around 0.05 by using the MDX-based procedure, while the chance of the same event is about 0.5 by using the MDR-based procedure. Thus, nearly half of the diagnoses cannot control MDP at the nominal level if the MDR-based procedure is used. In practice, it is nontrivial to adjust the nominal MDR level so that MDP can be controled properly. More numerical examples will be presented in later sections to compare the related procedures.

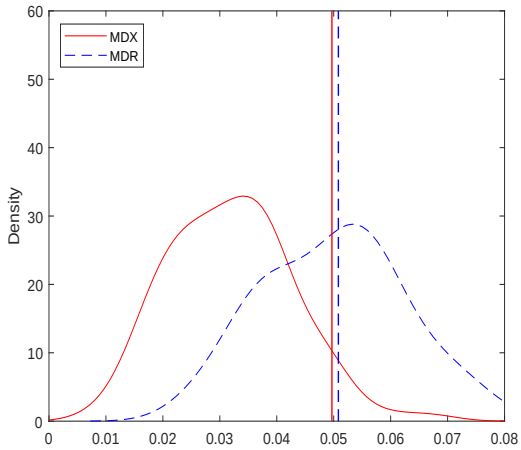


Figure 1: Density curves of MDP in the toy example for comparing the proposed MDX-based procedure with the MDR-based procedure by Li et al. (2020a). In this example, both $\alpha$ and $\tau$ are set to be 0.05.

The remainder of the paper is organized as follows. In Section 2, the problem of reliable fault diagnosis of HDS under the HMM dependence is formulated in a multiple-testing framework for controlling MDX. In Section 3, the oracle and data-driven diagnostic procedures are described. Some simulation results are given in Section 4 to evaluate the numerical performance of the proposed procedures in various cases considered. In Section 5, the data-driven diagnostic procedure is applied to a real-world example. In Section 6, we conclude the paper with several remarks.

# 2 A Reliable Fault Diagnosis Framework Using a Hidden Markov Model

Suppose there are $m$ data streams in a high-dimensional process considered. When the process is in-control (IC), the $m$-dimensional observation at time $t$ is denoted as $\boldsymbol{x}_t^{IC}$. Without loss of generality, it is assumed that the IC mean vector is $\mathbf{0}$. After an anomaly occurs, the process becomes OC. An online monitoring control chart can give us a signal of the process shift and a change-point detection approach can be used afterwards to estimate the specific time when the shift occurs (cf., Qiu 2014, Samuel and Pignatiello 2001). This

paper focuses on figuring out the OC data streams after a signal has been given by a control chart and the shift location has been estimated by a change-point detection approach. For simplicity of discussion, it is assumed here that the process shift has been signaled correctly by a SPC method (e.g., Zamba and Hawkins, 2006; Zou et al., 2011), and a small number of OC observations, denoted as $\{\boldsymbol{x}_j^{OC}, j = 1, ..., n\}$, is available for fault diagnosis. Note that these assumptions are made for both the proposed methods and all the alternative fault diagnosis methods considered in the paper, and thus they would not change the comparative results of the related methods. The OC process mean vector is denoted as $\boldsymbol{\mu}$. Apparently, some components of $\boldsymbol{\mu}$ would be non-zero due to the mean shift. In this paper, we further assume that the OC observations $\{\boldsymbol{x}_j^{OC}\}$ are independent over time for simplicity, but the $m$ components of a given observation can be mutually correlated. That is, in terms of spatial-temporal data analysis, we assume that observed data could be spatially correlated, but independent in the time domain. In practice, the temporal data correlation can be removed in advance by a time series model or another data decorrelation approach (e.g., Apley and Tsung, 2002, Qiu et al., 2020).

We use the sample mean of the OC observations $\boldsymbol{x}^{\boldsymbol{OC}} = \sum_{j=1}^{n} \boldsymbol{x}_j^{OC}/n$ to construct a fault diagnosis procedure. Hereafter, the notation $\boldsymbol{x}^{\boldsymbol{OC}}$ is simplified to $\boldsymbol{x} = (x_1, ..., x_m)^T$ for convenience. Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)^T$ be the unknown hidden status of $\boldsymbol{x}$, where $\theta_i = 1$ means that the $i$th data stream is OC while $\theta_i = 0$ means IC. In many real applications, there might be correlation among the hidden statuses of different data streams, which is highly informative but often ignored in the SPC research. In the literature, HMM is a commonly-used model to describe such a dependence structure (Ephraim and Merhav, 2002; Sun and Cai, 2009), which is briefly described below. Assume that the unobservable $\theta_i$s form a stationary Markov chain. Specifically, let $\boldsymbol{A} = (a_{jk})_{2\times 2}$ be the transition matrix, where $a_{jk} = P(\theta_i = k|\theta_{i-1} = j)$, for $j, k = 0, 1$, are the transition probabilities that do not depend on $i$ and have the standard constraints that $0 < a_{jk} < 1$ and $a_{j0} + a_{j1} = 1$. If a positive correlation exists among the hidden statuses, then the OC data streams would tend to appear in clumps, which should be a natural feature of many high-dimensional applications. Then, the observed data $\boldsymbol{x}$ can be generated from the following conditional distribution:

$$x_i|\theta_i \sim (1 - \theta_i)F_{0i} + \theta_i F_{1i}, \quad i = 1, ..., m, \tag{1}$$

where $F_{0i}$ and $F_{1i}$ are respectively the IC and OC distributions of each random variable. $F_{0i}$s are generally assumed known. Since normal inverse transformation $\Phi^{-1}(F_{0i}(\cdot))$ can be applied, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal c.d.f., it is reasonable to assume that $F_{0i}$s are standard normal. $F_{1i}$s suffer from mean shifts, and the shift sizes come from certain prior distribution. To facilitate the derivation of the proposed procedures, we further assume that the distribution of the mean shifts is a degenerate distribution, i.e., $F_{1i} = F_1$. Such a mixture model and its variations have been used widely in high-dimensional data analysis (cf., e.g., Sun and Cai, 2007, 2009; Zou et al., 2015; Li et al., 2020a). Because $\boldsymbol{x}$ is actually the sample mean of $n$ OC observations, as long as $n$ is reasonably large, we can assume based on the central limit theory that $F_1$ is asymptotically normal.

It has been demonstrated in the literature that high-dimensional fault diagnosis (i.e. find the OC data streams with non-zero $\theta_i$s) can be formulated properly as a large-scale multiple-testing problem in which the following $m$ hypotheses are tested simultaneously (Li et al., 2020a):

$$H_i^0 : \theta_i = 0 \quad \text{versus} \quad H_i^1 : \theta_i = 1, \quad i = 1, ..., m. \tag{2}$$

The decision rule of the above hypothesis testing problem, denoted as $\boldsymbol{\delta} = (\delta_1, ..., \delta_m)^T \in \{0, 1\}^m$, generally has the form

$$\delta_i = I[T_i(\boldsymbol{x}) < c], \tag{3}$$

where $\delta_i$ equals 0 if we claim that the $i$th data stream is IC and 1 otherwise, $T_i(\boldsymbol{x})$ is a test statistic that maps $\boldsymbol{x}$ to a real value, $c$ is a universal threshold, and $I[\cdot]$ is the indicator function. Different choices of $T_i(\boldsymbol{x})$ will lead to different diagnostic performance, which will be discussed in detail in later sections. The performance of a single diagnosis can be evaluated by the missed discovery proportion (MDP), which is defined to be the proportion of false negatives out of all OC data streams as follows:

$$\text{MDP} = \frac{\sum_{i=1}^{m} \theta_i(1 - \delta_i)}{(\sum_{i=1}^{m} \theta_i) \vee 1}.$$

Note that MDP is calculated based on a single diagnosis, which is highly variable. Therefore, we suggest using the missed discovery exceedance MDX as the performance measure. Specifically, let $\tau \in (0,1)$ be a prespecified tolerance level. MDX at level $\tau$ can then be defined as $\text{MDX}_\tau = P(\text{MDP} > \tau)$, i.e., the probability that MDP exceeds $\tau$. By controlling $\text{MDX}_\tau$ at a desirable level, the issue of high unreliability of MDP can be avoided. In this paper, controlling $\text{MDX}_\tau$ at a desirable level is referred to as MDX-control. Obviously, MDX-control has taken into account the large variability of MDP, which is desirable for a reliable fault diagnosis of HDS. By the way, the idea to use the exceedance probability can also be found in the literature when handling the large variability of the false discovery proportion (FDP). See, for example, Pacifico et al. (2004) and Genovese and Wasserman (2006).

In practice, it is generally difficult to calculate $\text{MDX}_\tau$. The main reason is that the distribution of MDP is very complicated, and its tail probability needs to be calculated through a complex integral involving $\boldsymbol{x}$ and $\boldsymbol{\theta}$, which is often difficult in practice since the sample mean of the OC observations only has a single vector value for a given diagnosis. To overcome this difficulty, we introduce a useful alternative to $\text{MDX}_\tau$ by considering the following conditional probability:

$$\text{MDX}_{\tau,\boldsymbol{x}} = P(\text{MDP} > \tau | \boldsymbol{x}).$$

It can be proved that if $\text{MDX}_{\tau,\boldsymbol{x}} = \alpha$ for all possible values of $\boldsymbol{x}$, then $\text{MDX}_\tau = \alpha$. In the rest of this paper, we will derive a reliable diagnostic procedure by controlling the value of $\text{MDX}_{\tau,\boldsymbol{x}}$ at a desired level $\alpha$.

# 3 New Statistical Methodology

Given the multiple-testing-based fault diagnosis framework discussed in Section 2, we describe our proposed diagnostic procedures in this section for correlated HDS. First, let us further discuss about the test statistic $T_i(\boldsymbol{x})$ in (3). In order to properly handle between-stream data dependence, we introduce the local index of significance (LIS) for the $i$th data

stream by

$$\mathrm{LIS}_i = P(\theta_i = 0|\boldsymbol{x}), \tag{4}$$

which is the probability that the $i$th stream is IC given $\boldsymbol{x}$ (cf., Sun and Cai, 2009). Intuitively, $\mathrm{LIS}_i$ would be small when the $i$th data stream is OC and vice versa. Note that when different data streams are independent of each other, $\mathrm{LIS}_i$ only depends on $x_i$ and reduces to the local false discovery rate (Lfdr) proposed by Efron (2004). Sun and Cai (2007) showed that a multiple-testing procedure for FDR-control is optimal in the independent case by using Lfdr. Nevertheless, LIS is much more appropriate than Lfdr when the data streams are correlated. The main reason is that the OC streams tend to appear in clusters, and one should treat a data stream surrounded by OC streams differently from the ones surrounded by IC streams. In such cases, LIS is asymmetric in the sense that pooling information from adjacent data streams can improve decision-making. By contrast, procedures that threshold Lfdr are symmetric rules under which the data streams are exchangeable, making them undesirable to handle correlated data streams. Therefore, we suggest thresholding LIS in our diagnostic procedures. In Section 3.1, an oracle diagnostic procedure that uses LISs as test statistics is developed. Note that "oracle" here means the situation that all distributional information is assumed completely known. In such a case, the decision rule has the form

$$\delta_i = I(\mathrm{LIS}_i < c). \tag{5}$$

In Section 3.2, a data-driven diagnostic procedure that mimics the oracle version is proposed for practical use. The advantages of thresholding LISs over other alternative test statistics are demonstrated numerically in Section 4.

## 3.1 Oracle Diagnostic Procedure

Based on the definitions of $\delta_i$s and $\mathrm{MDX}_{\tau,\boldsymbol{x}}$, it can be seen that $\mathrm{MDX}_{\tau,\boldsymbol{x}}$ is a function of the threshold value $c$. Therefore, in order to control $\mathrm{MDX}_{\tau,\boldsymbol{x}}$ at the level $\alpha$, $c$ must be chosen carefully. To this end, we suggest using the following two-step approach: (i) calculate the value of $\mathrm{MDX}_{\tau,\boldsymbol{x}}$ for a given $c$, and (ii) search the value of $c$ such that $\mathrm{MDX}_{\tau,\boldsymbol{x}}$ reaches the

level $\alpha$. These two steps are discussed in the two parts below.

**Calculation of $\mathbf{MDX}_{\tau,\mathbf{x}}$.** We propose an efficient approach to calculate the value of $\text{MDX}_{\tau,\boldsymbol{x}}$ for a given value of $c$. After some simple algebraic manipulations, it can be checked that $\text{MDX}_{\tau,\boldsymbol{x}}$ can be written as

$$\begin{aligned}
\text{MDX}_{\tau,\boldsymbol{x}} &= P(\text{MDP} > \tau|\boldsymbol{x}) = E[I(\text{MDP} > \tau)|\boldsymbol{x}] \\
&= \sum_{\boldsymbol{\theta}} I_{(0,+\infty)}(\text{MDP} - \tau) f(\boldsymbol{\theta}|\boldsymbol{x}) \\
&= \sum_{\boldsymbol{\theta}} I_{(0,+\infty)}(\text{MDP} - \tau) \frac{f(\theta_m, \theta_{m-1}, ..., \theta_1, \boldsymbol{x})}{f(\theta_{m-1}, \theta_{m-2}, ..., \theta_1, \boldsymbol{x})} \frac{f(\theta_{m-1}, \theta_{m-2}, ..., \theta_1, \boldsymbol{x})}{f(\theta_{m-2}, \theta_{m-3}, ..., \theta_1, \boldsymbol{x})} \cdots \frac{f(\theta_1, \boldsymbol{x})}{f(\boldsymbol{x})} \\
&= \sum_{\boldsymbol{\theta}} I_{(0,+\infty)}(\text{MDP} - \tau) f(\theta_m|\theta_{m-1}, ..., \theta_1, \boldsymbol{x}) f(\theta_{m-1}|\theta_{m-2}, ..., \theta_1, \boldsymbol{x}) \cdots f(\theta_1|\boldsymbol{x}), \quad (6)
\end{aligned}$$

where $f(\cdot)$ and $f(\cdot|\cdot)$ stand for the density function and the probability mass function (p.m.f.), respectively. From Equation (6), it can be seen that the computational complexity for calculating $\text{MDX}_{\tau,\boldsymbol{x}}$ is $O(2^m)$, which is too intensive for real applications even when the distributional information is assumed known. To overcome this difficulty, we suggest using the Monte Carlo method to first generate a sufficiently large number of random vectors and then calculate the corresponding values of MDP. Then, the value of $\text{MDX}_{\tau,\boldsymbol{x}}$ can be approximated by $\sum_{j=1}^{N} I[\text{MDP}_j > \tau]/N$, where $\text{MDP}_j$ is the MDP value calculated from the $j$th random vector, and $N$ is the number of random vectors generated. Such a Monte Carlo approach has been used widely in the literature for obtaining numerical solutions to problems that are too complicated to solve analytically.

After an observation $\boldsymbol{x}$ is collected, we can first determine the p.m.f. of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ as follows. By Equation (6), this conditional p.m.f. can be written as

$$f(\boldsymbol{\theta}|\boldsymbol{x}) = f(\theta_1|\boldsymbol{x}) \prod_{i=2}^{m} f(\theta_i|\theta_{i-1}, ..., \theta_1, \boldsymbol{x}). \quad (7)$$

This motivates us the following recursive computation of $f(\boldsymbol{\theta}|\boldsymbol{x})$: $f(\theta_1|\boldsymbol{x})$ can be computed first, then $f(\theta_2|\theta_1, \boldsymbol{x})$ can be computed, and so forth. It can be shown easily that the distribution of $\theta_i$ given $\theta_{i-1}, ..., \theta_1$ and $\boldsymbol{x}$ is $Bernoulli(f(\theta_i = 1|\theta_{i-1}, ..., \theta_1, \boldsymbol{x}))$. The calculation of $f(\theta_i = 1|\theta_{i-1}, ..., \theta_1, \boldsymbol{x})$ can be further divided into calculating $f(\theta_i, ..., \theta_1, \boldsymbol{x})$ and

$f(\theta_{i-1}, ..., \theta_1, \boldsymbol{x})$ based on the following conditional probability formula:

$$f(\theta_i|\theta_{i-1}, ..., \theta_1, \boldsymbol{x}) = \frac{f(\theta_i, ..., \theta_1, \boldsymbol{x})}{f(\theta_{i-1}, ..., \theta_1, \boldsymbol{x})}.$$

After some simple manipulations, $f(\theta_i, ..., \theta_1, \boldsymbol{x})$ can be written as

$$\begin{aligned}
f(\theta_i, ..., \theta_1, \boldsymbol{x}) &= \sum_{\theta_{i+1},...,\theta_m=0}^{1} \prod_{k=1}^{m} [\theta_k f_1(x_k) + (1-\theta_k) f_0(x_k)] a_{\theta_{k-1}\theta_k} \\
&= \left\{ \prod_{k=1}^{i} [\theta_k f_1(x_k) + (1-\theta_k) f_0(x_k)] a_{\theta_{k-1}\theta_k} \right\} \\
&\quad \cdot \sum_{\theta_{i+1},...,\theta_m=0}^{1} \prod_{k=i+1}^{m} [\theta_k f_1(x_k) + (1-\theta_k) f_0(x_k)] a_{\theta_{k-1}\theta_k} \\
&= \left\{ \prod_{k=1}^{i} [\theta_k f_1(x_k) + (1-\theta_k) f_0(x_k)] a_{\theta_{k-1}\theta_k} \right\} f(x_{i+1}, ..., x_m|\theta_i) \\
&\triangleq \left\{ \prod_{k=1}^{i} [\theta_k f_1(x_k) + (1-\theta_k) f_0(x_k)] a_{\theta_{k-1}\theta_k} \right\} \beta_i(\theta_i), \quad i = 1, ..., m,
\end{aligned}$$

where $\beta_i(\theta_i) = f(x_{i+1}, ..., x_m|\theta_i)$. Consequently, $f(\theta_i|\theta_{i-1}, ..., \theta_1, \boldsymbol{x})$ can be written as:

$$f(\theta_i = j|\theta_{i-1}, ..., \theta_1, \boldsymbol{x}) = f_j(x_i) a_{\theta_{i-1},\theta_i=j} \frac{\beta_i(j)}{\beta_{i-1}(\theta_{i-1})}, \quad j = 0, 1. \tag{8}$$

Given Equation (8), we can now generate a group of $\theta_i$s based on their conditional distributions. In order to calculate MDP, we still need to calculate $\text{LIS}_i$ defined in (4). To this end, $\text{LIS}_i$ can be written as

$$\begin{aligned}
\text{LIS}_i &= P(\theta_i = 0|\boldsymbol{x}) = \frac{f(\boldsymbol{x}, \theta_i = 0)}{f(\boldsymbol{x})} \\
&= \frac{f(x_1, ..., x_i, \theta_i = 0) f(x_{i+1}, ..., x_m|\theta_i = 0)}{f(\boldsymbol{x})} \\
&\triangleq \frac{\eta_i(0)\beta_i(0)}{\eta_i(0)\beta_i(0) + \eta_i(1)\beta_i(1)},
\end{aligned} \tag{9}$$

where $\eta_i(\theta_i) = f(x_1, ..., x_i, \theta_i)$. In Equations (8) and (9), $\eta_i(\theta_i)$ and $\beta_i(\theta_i)$ both need to be calculated. This task can be accomplished recursively by using the forward-backward algorithm. To be more specific, after initializing $\eta_1(j) = P(\theta_1 = j) f_j(x_1)$ and $\beta_m(j) = 1$, we

Table 1: Monte-Carlo calculation of $\text{MDX}_{\tau,\boldsymbol{x}}$.

**Preparation Step**:
Collect the observed data $\boldsymbol{x}$.

**Generation Step**:
1. Generate a group of $N$ $\boldsymbol{\theta}$ values conditionally on $\boldsymbol{x}$ ($N$ is chosen to be $10^4$ in this paper).
2. According to the decision rules $\delta_i = I(\text{LIS}_i < c)$, calculate $R_j = I(\text{MDP} > \tau), j = 1, ..., N$.

**Calculation Step**:
Calculate the approximate value of $\text{MDX}_{\tau,\boldsymbol{x}}$ by $\sum_{j=1}^{N} R_j/N$.

can use the following recursive formulas:

$$\eta_{i+1}(j) = \left\{ \sum_{k=0}^{1} \eta_i(k) a_{kj} \right\} f_j(x_{i+1}),$$

and

$$\beta_i(j) = \sum_{k=0}^{1} a_{jk} f_k(x_{i+1}) \beta_{i+1}(k), \quad j = 0, 1.$$

More details can be found in Sun and Cai (2009).

After the LIS values are calculated, we can calculate MDP for a given $c$, and then approximate the $\text{MDX}_{\tau,\boldsymbol{x}}$ value by using the Monte Carlo approach. The entire Monte-Carlo-based calculation process is summarized in Table 1.

**Searching for the optimal threshold** $c$**.** After calculating the approximate value of $\text{MDX}_{\tau,\boldsymbol{x}}$ for a given $c$, in this part we aim to find the optimal value of $c$ so that $\text{MDX}_{\tau,\boldsymbol{x}}$ can be controlled at the nominal level $\alpha$. To this end, the following proposition confirms the existence of such an optimal value theoretically.

**Proposition 1.** *Consider the mixture model (1) under the HMM dependence and the decision rules $\delta_i$ defined in (5). Let $Q(c)$ denote the value of $MDX_{\tau,\boldsymbol{x}}$ with the threshold $c$. Then, for given $\tau \in (0,1)$ and $\alpha \in (0,1)$, we have the following results:*

*(i) $Q(c)$ is non-increasing in $c$, and*
*(ii) $Q(c^*) = \alpha$, where*

$$c^* = \inf\{c : Q(c) \leq \alpha\}.$$

The proof of Proposition 1 is quite straightforward. Specifically, MDP can be written as

$$\text{MDP} = \frac{\sum_{i=1}^{m} \theta_i I(\text{LIS}_i \geq c)}{(\sum_{i=1}^{m} \theta_i) \vee 1},$$

which is non-increasing in $c$. Thus, $\text{MDX}_{\tau,\boldsymbol{x}} = P(\text{MDP} > \tau|\boldsymbol{x})$ is non-increasing in $c$ as well, making the existence and uniqueness of $c^*$ valid. By Proposition 1, we can search for the optimal threshold $c^*$ numerically by using the bisection searching algorithm, which is described below:

Step 1 Based on the proposed Monte-Carlo approach, we first find $c_1$ and $c_2$ such that $Q(c_1) < \alpha$ and $Q(c_2) > \alpha$.

Step 2 Calculate $Q(c_3)$, where $c_3 = (c_1 + c_2)/2$.

Step 3 If $Q(c_3) > \alpha$, then assign $c_2 = c_3$; If $Q(c_3) < \alpha$, then assign $c_1 = c_3$.

Step 4 Repeat Steps 2 and 3 until $Q(c_3)$ is sufficiently close to $\alpha$. Then, the last $c_3$ value is regarded as the searched value of $c^*$.

By now, we have defined all components of the oracle diagnostic procedure for MDX-control, which is summarized below:

Let $\gamma = \{i : \text{LIS}_i < c^*\}$. Then, reject $H_i^0$ and claim that the $i$th data stream is OC if $i \in \gamma$.

## 3.2 Data-Driven Diagnostic Procedure

In the oracle diagnostic procedure discussed in the previous subsection, all the distributional information is assumed known, which is invalid in practice. To overcome this difficulty,

we propose a data-driven diagnostic procedure in this subsection for practical purposes in cases when the underlying distribution is unknown.

We first need to estimate the unknown HMM parameters. Let $\vartheta = (\{p_0, p_1\}, \{F_0, F_1\}, \boldsymbol{A})$ be the collection of the HMM parameters, where $p_j = P(\theta_i = j)$ denotes the asymptotic stationary distribution of $\theta_i$ since $\frac{1}{m}\sum_{i=1}^{m} I(\theta_i = j) \to p_j$, for $j = 0, 1$, by the convergence theorem of a Markov chain (cf., Durrett, 2005). In the literature, the maximum likelihood estimate (MLE) has been widely used for estimating $\vartheta$ (Leroux, 1992; Bickel et al., 1998). By MLE, the parameters are estimated by maximizing the likelihood function. In the current problem, the likelihood function has the expression

$$L(\vartheta; \boldsymbol{x}, \boldsymbol{\theta}) = p_{\theta_1} \prod_{i=2}^{m} a_{\theta_{i-1}\theta_i} \prod_{i=1}^{m} f_{\theta_i}(x_i).$$

Note that $\boldsymbol{x}$ is the sample mean of $n$ OC observations. So, it is reasonable to assume based on the central limit theorem that $f_0(\cdot)$ and $f_1(\cdot)$ are the density functions of $N(\mu_0, \sigma^2)$ and $N(\mu_1, \sigma^2)$, respectively, where $\sigma^2$ can be estimated by the sample variance. Let $\widehat{\vartheta}$ denote the MLE of $\vartheta$. Then, it has been well demonstrated in the literature that $\widehat{\vartheta}$ is consistent and asymptotically normal under some regularity conditions. The MLE $\widehat{\vartheta}$ can be obtained by using a numerical optimization algorithm, such as the well-known EM algorithm and the gradient search algorithm. The EM algorithm for obtaining the MLE $\widehat{\vartheta}$ in our setting is summarized in Table 2.

We now plug-in $\widehat{\vartheta}$ to obtain the plug-in statistics $\widehat{f}(\theta_i|\theta_{i-1}, ..., \theta_1, \boldsymbol{x})$ and $\widehat{\text{LIS}}_i$, which can be computed by using the forward-backward algorithm described in Section 3.1. Then, in light of the oracle diagnostic procedure, the Monte Carlo method and the bisection search algorithm in Section 3.1 can also be implemented to find the optimal threshold value numerically, denoted as $c^{**}$. Finally, the proposed data-driven diagnostic procedure can be summarized below:

Let $\gamma^* = \{i : \widehat{\text{LIS}}_i < c^{**}\}$. Then, reject $H_i^0$ and claim that the $i$th data stream is OC if $i \in \gamma^*$.

At the end of this section, we would like to point out that the proposed data-driven diagnostic procedure is described above when the shift magnitudes of different OC data

Table 2: The EM algorithm for estimating the HMM parameters.

**Preparation Step**:
Give initial values for the HMM parameters: $p_j^{(0)}$, $a_{jk}^{(0)}$, $\mu_0^{(0)}$, $\mu_1^{(0)}$. Set $t = 0$.

**E Step**:
Compute the following quantities:

- $\eta_i^{(t)}(j) = f(x_1, ..., x_i, \theta_i = j)$
- $\beta_i^{(t)}(j) = f(x_{i+1}, ..., x_m | \theta_i = j)$
- $\gamma_i^{(t)}(j) = \eta_i^{(t)}(j)\beta_i^{(t)}(j)/[\eta_i^{(t)}(0)\beta_i^{(t)}(0) + \eta_i^{(t)}(1)\beta_i^{(t)}(1)]$
- $\xi_i^{(t)}(j,k) = P(\theta_i = j, \theta_{i+1} = k|\boldsymbol{x}) = \gamma_i^{(t)}(j)a_{jk}^{(t)}f_k(x_{i+1})\beta_{i+1}^{(t)}(k)/\beta_i^{(t)}(j)$

**M Step**:
Set $t = t + 1$ and update the parameters:

- $p_j^{(t)} = \gamma_1^{(t-1)}(j)$
- $a_{jk}^{(t)} = \sum_{i=1}^{m-1} \xi_i^{(t-1)}(j,k) / \sum_{i=1}^{m-1} \gamma_i^{(t-1)}(j)$
- $\mu_0^{(t)} = [\sum_{i=1}^{m} \gamma_i^{(t-1)}(0)x_i] / \sum_{i=1}^{m} \gamma_i^{(t-1)}(0)$
- $\mu_1^{(t)} = [\sum_{i=1}^{m} \gamma_i^{(t-1)}(1)x_i] / \sum_{i=1}^{m} \gamma_i^{(t-1)}(1)$

**Iterate the E Step and the M Step until convergence.**

streams are assumed to be the same. This assumption, however, can be lifted easily, by changing $F_1$ to a normal mixture $\sum_{j=1}^{J} w_j N(\mu_j^{OC}, \sigma^2)$ with $\sum_{j=1}^{J} w_j = 1$, where $\{\mu_j^{OC}, j = 1, 2, \ldots, J\}$ denote the OC means of the $J$ OC data streams, and $\{w_j, j = 1, 2, \ldots, J\}$ are the weights. In such cases, the whole normal mixture is regarded as a single OC state, and the parameters (i.e., $\{\mu_j^{OC}, w_j\}$) can be estimated by the EM algorithm efficiently. See Sun and Cai (2009) for a related discussion.

# 4 Simulation Studies

In this section, we investigate systematically the numerical performance of the proposed oracle and data-driven diagnostic procedures. The section is organized in three parts. In Section 4.1, we study the impact of using different test statistics $T_i(\boldsymbol{x})$. In Section 4.2, we compare the proposed diagnostic procedures with the diagnostic procedure proposed by Li et al. (2020a) which was shown to have a better performance than its rivals in the SPC literature. In Section 4.3, we will confirm that the diagnostic performance of the oracle procedure can be attained asymptotically by the data-driven procedure, which ensures the latter procedure to be feasible for practical use. In Section 4.4, we study the robustness of the proposed data-driven procedure when there is a false alarm.

## 4.1 Impact of Different Test Statistics

In this part, we focus on the proposed oracle procedure, and study its diagnostic performance with different test statistics. Specifically, besides LIS, two other test statistics are considered here: Lfdr and $p$-value. When determining the significance level for a data stream, an Lfdr or $p$-value approach would consider each data stream separately. One can expect that the approaches using Lfdr and $p$-value would be less efficient than the one using LIS, since the former ones do not take into account the between-stream correlation.

In all simulation examples, we choose the data dimension to be $m = 1000$. The Markov chain $\boldsymbol{\theta}$ is generated with the initial state 0 and the transition matrix $\boldsymbol{A} = (a_{jk})_{2\times 2}, j, k = 0, 1$, where $\boldsymbol{A}$ is chosen to be one of the following three matrices:

- Independent case: $a_{00} = 0.5, a_{01} = 0.5, a_{10} = 0.5, a_{11} = 0.5$.
- Weak correlation case: $a_{00} = 0.6, a_{01} = 0.4, a_{10} = 0.35, a_{11} = 0.65$.
- Strong correlation case: $a_{00} = 0.9, a_{01} = 0.1, a_{10} = 0.2, a_{11} = 0.8$.

The steady-state OC probability is $p_1 = a_{01}/(a_{01} + a_{10})$, and the values of $p_1$ in the above three cases are 0.5, 0.53 and 0.33, respectively. The observed distribution of $x_i$ given $\theta_i$ is assumed to be $(1 - \theta_i)N(0, 1) + \theta_i N(\mu, 1)$, where $\mu$ is the OC mean. For each simulation,

we first generate 10,000 groups of $\boldsymbol{\theta}$ and $\boldsymbol{x}$ with a given HMM model and the observed distribution specified above. Then, the proposed oracle diagnostic procedure runs 10,000 times to calculate the actual MDP values. The diagnostic results of the proposed oracle procedure using LIS, Lfdr and $p$ values when $\mu = 2$ are summarized in Figure 2, in which the density functions of MDP and the corresponding expected numbers of false positives (EFP), defined as EFP $= E[\sum_{i=1}^{m}(1-\theta_i)\delta_i]$, are plotted. The vertical lines in the density plots represent the $100(1-\alpha)$th percentiles of MDP. It is important to note that while controlling $\text{MDX}_\tau$ at level $\alpha$, a method with a smaller EFP value would be more preferable in practice, as a smaller EFP value means less false positives. The nominal MDX level $\alpha$ and the threshold value $\tau$ are both set at 0.05 for convenience. The simulation results of other choices of $\alpha$ and $\tau$ are similar, and thus are omitted here.

From the plots in Figure 2, it can be observed that $\text{MDX}_\tau$ can be properly controlled under $\alpha$ by using all three test statistics, but the EFP values of the method using the $p$-value are significantly larger than those of the method using the other two statistics. Also, in the independent case, the performance of LIS and Lfdr is almost identical. This should not be surprising since LIS reduces to Lfdr when the data streams are independent of each other. As the between-stream correlation gets stronger, the advantage of LIS over Lfdr becomes more and more obvious. This finding implies that, if the between-stream correlation can be modelled appropriately, it can be a blessing for a reliable fault diagnosis (i.e., EFP would decrease when $a_{11}$ increases); but, if it is ignored, the between-stream correlation can cause a problem. Therefore, we can conclude from this example that thresholding LIS is a reasonable choice for MDX-control when between-stream correlation exists.

We also investigate the diagnostic performance under the three test statistics with various different model parameter settings, including: (i) different shift sizes when $\mu$ changing among 1, 1.5 and 2; (ii) different numbers of variables when $m$ changing among 1,000, 2,000 and 3,000; and (iii) different steady-state OC probabilities when

$$\boldsymbol{A} = \begin{pmatrix} 0.95, & 0.05 \\ 0.05(0.8+\xi)/(0.2-\xi), & 1-0.05(0.8+\xi)/(0.2-\xi) \end{pmatrix},$$

and $\xi$ changing among 0.05, 0.1 and 0.15. For cases (i) and (ii), only the strong correlation case is considered for simplicity. In case (iii), the steady-state OC probabilities corresponding

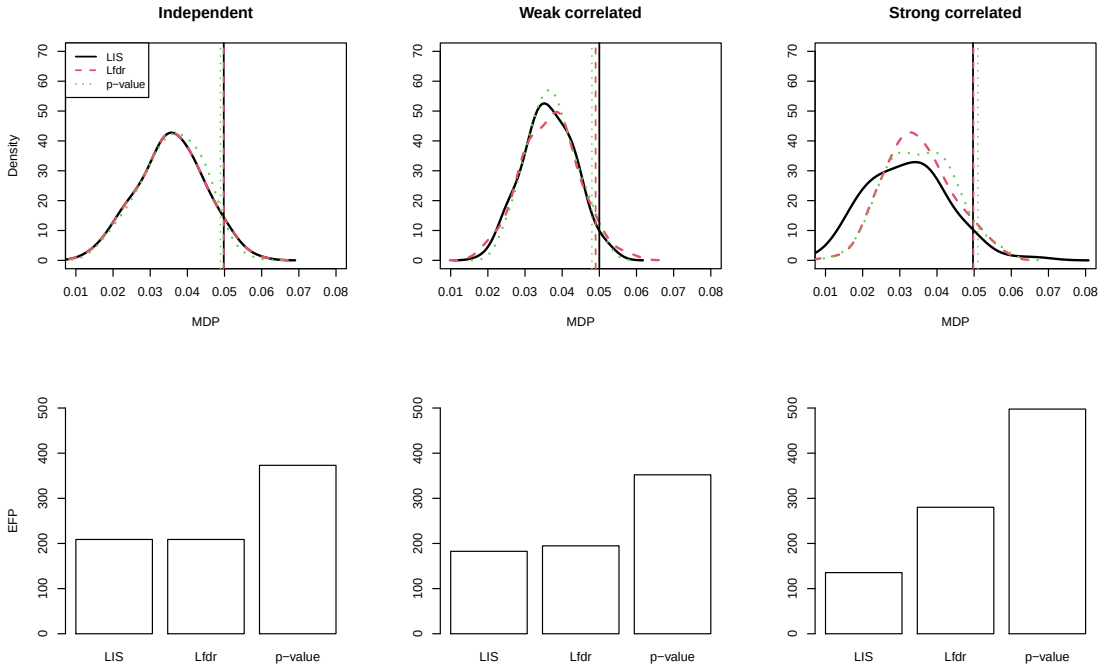


Figure 2: Diagnostic performance of the oracle procedure for MDX-control by using the three thresholding methods LIS, Lfdr, and $p$-value. The first row shows the density curves of MDP and the second row shows the EFP levels. The vertical lines denote the $100(1-\alpha)$th percentiles of MDP of the procedures for MDX-control.

to the three values of $\xi$ are 0.15, 0.1 and 0.05, respectively. To save space in the main paper, the simulation results in these three cases are shown in Figures 7-9 in the Supplementary File. From the figures, we can make similar conclusions as those from Figure 2 that $\text{MDX}_\tau$ is controlled well in our proposed method in all cases considered and that thresholding LIS would be better than the other two thresholding approaches.

## 4.2 Comparison Between MDX-control and Some Existing Methods

In this part, we compare the proposed diagnostic procedure for MDX-control that uses the LIS thresholding approach with some existing representative diagnostic methods in the literature. One existing method considered is the diagnostic procedure proposed by Li et al. (2020a) that controls MDR. MDR is the mathematical expectation of MDP, and the

MDR-based procedure tries to minimize EFP while controlling MDR at level $\alpha$. Another existing method considered is the LASSO-based multivariate diagnostic framework (denoted as LASSO) suggested in Zou et al. (2011). It should be noted that the hype-parameters in LASSO are tuned such that its nominal MDX is guaranteed. Figure 3 presents the diagnostic performance of MDX-control and the above-mentioned two existing methods in the strong correlation case. In the three plots of the upper row in the figure, the vertical dashed lines denote the 50th percentiles (i.e., medians) of the MDP values of the procedure for MDR-control, and the vertical solid and dotted lines are the $100(1-\alpha)$th percentiles of the MDP values of the proposed procedure for MDX-control and the procedure LASSO, respectively.

From the plots, it can be seen that the probability for MDP being larger than $\tau$ can be controlled properly by controlling the MDX at level $\alpha$. As a comparison, by using the procedure for MDR-control, even in cases when MDR can be controlled at $\alpha$, it is likely to result in a diagnosis with MDP being significantly larger than $\alpha$, as evidenced by the symmetrical structure of the density curves of MDP for the MDR-control procedure where the approximate medians are close to $\alpha$. In addition, given the fact that MDX-control would miss less OC data streams than MDR-control, the numbers of IC data streams that are diagnosed as OC for the MDX-control procedure would also be smaller than those by the MDR-control procedure in all three cases considered. As for the LASSO procedure, it can be observed from the figure that even when the MDX is controlled at level $\alpha$, its EFP values are significantly larger than those of the MDX-control procedure, and the difference between their EFP values gets bigger as $\mu$ increases. Results in this example imply that the proposed diagnostic procedure for MDX-control would be more reliable and effective than the MDR-control and LASSO procedures in cases when the data streams are correlated.

## 4.3 Comparison Between the Proposed Oracle and Data-driven Procedures

In this part, we compare the proposed data-driven diagnostic procedure for MDX-control with the oracle version. We set $\mu$ at 0.5, and consider the strong correlation case considered in Section 4.1. In order to study thoroughly the robustness and effectiveness of these procedures, we consider three scenarios: (a) fix $\alpha$ and $\tau$, and study the diagnostic performance for various

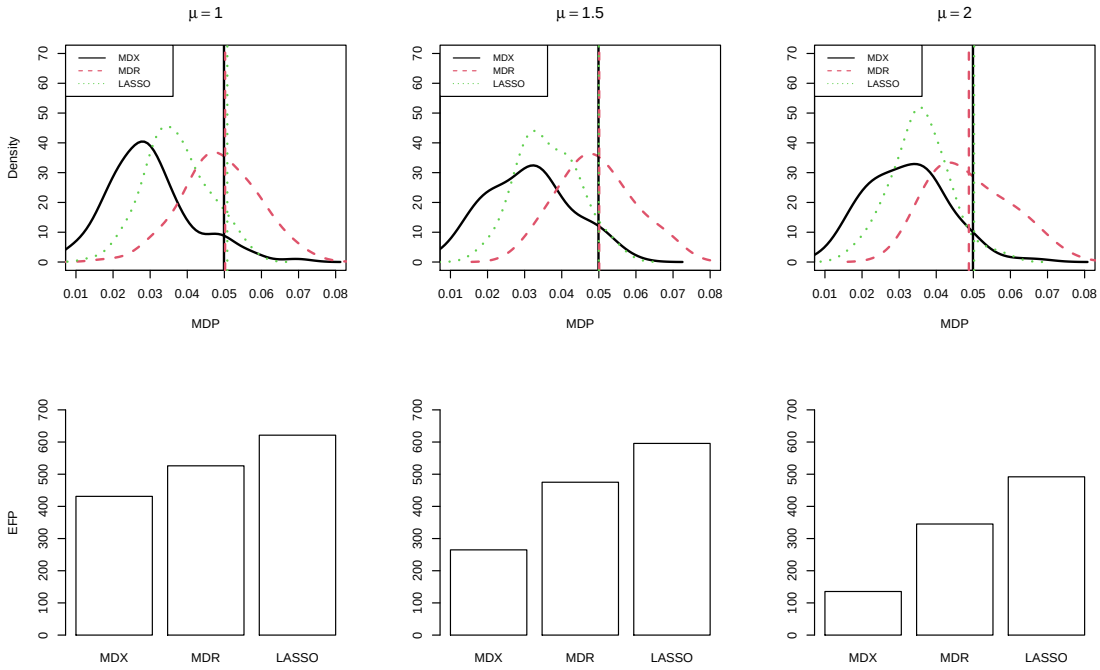


Figure 3: Diagnostic performance of the MDX-control, MDR-control and LASSO procedures. The first row shows the density curves of MDP and the second row shows the EFP levels. The vertical solid lines denote the $100(1-\alpha)$th percentiles of MDP of the MDX-control procedure, the vertical dashed lines are the medians of MDP of the MDR-control procedure, and the vertical dotted lines denote the $100(1-\alpha)$th percentiles of MDP of the LASSO procedure.

choices of $n$; (b) fix $n$ and $\tau$, and study the diagnostic performance for various choices of $\alpha$; and (c) fix $n$ and $\alpha$, and study the diagnostic performance for various choices of $\tau$. Other setups are kept to be the same as before. The simulation results are shown in Figure 4, where the MDX and EFP values are shown as functions of $n$, $\tau$ and $\alpha$ in respective scenarios. From the plots, it can be seen that the MDX levels of the oracle procedure are controlled well at the nominal level $\alpha$ in all cases considered and the lines of the oracle and data-driven procedures are almost identical, which indicates that the diagnostic performance of the proposed oracle procedure for MDX-control can be asymptotically attained by the data-driven procedure, and that the reliability and effectiveness of the data-driven procedure for MDX-control can be guaranteed in real-data applications. Furthermore, it can be noticed that the EFP level decreases as the two hype-parameters $\tau$ and $\alpha$ increases, which implies that our method tends to be more conservative when $\tau$ and $\alpha$ are smaller.

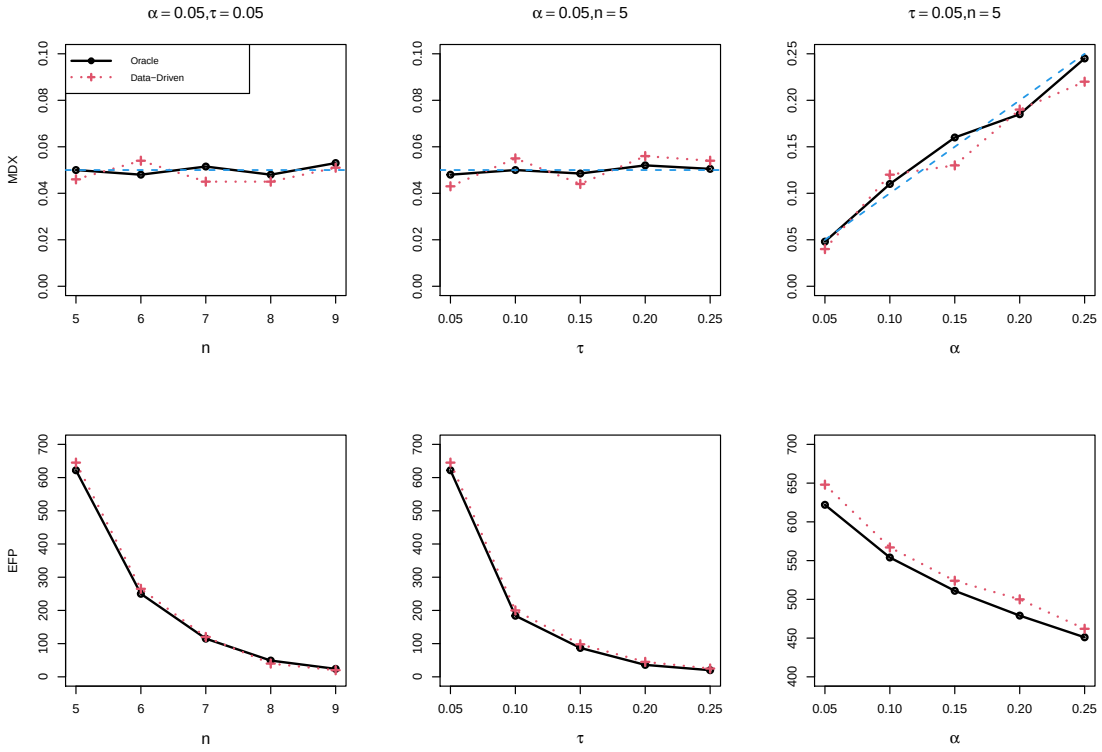


Figure 4: Comparison of the proposed oracle and data-driven diagnostic procedures for MDX-control. The first row shows the MDX levels and the second row shows the EFP levels. The blue dashed lines in the plots of the first row denote the nominal MDX levels.

## 4.4 Sensitivity Analysis

In this subsection, we conduct some sensitivity analysis to demonstrate the robustness and effectiveness of the proposed data-driven diagnostic procedure for MDX-control in cases with various invalid model assumptions.

**Diagnostic Performance Under Model Misspecifications.** In some applications, the model assumptions required by the proposed methodology may not be valid. To investigate the robustness of our procedure, we perform the following simulation studied in cases when the model is misspecified. First, we study the sensitivity of our method to the HMM assumption. To this end, the OC observations are generated from a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the $(i, j)$th element of $\boldsymbol{\Sigma}$ is set to be $0.5^{|i-j|}$, for all $i$ and $j$, all non-zero values in $\boldsymbol{\mu}$ are set to be 2, each element of $\boldsymbol{\mu}$ has 50% chance to be 0 or 1, and all elements of $\boldsymbol{\mu}$ are independently generated in each simulation. The other settings are the same as that in Figure 2. The simulation results of the proposed data-driven procedure are

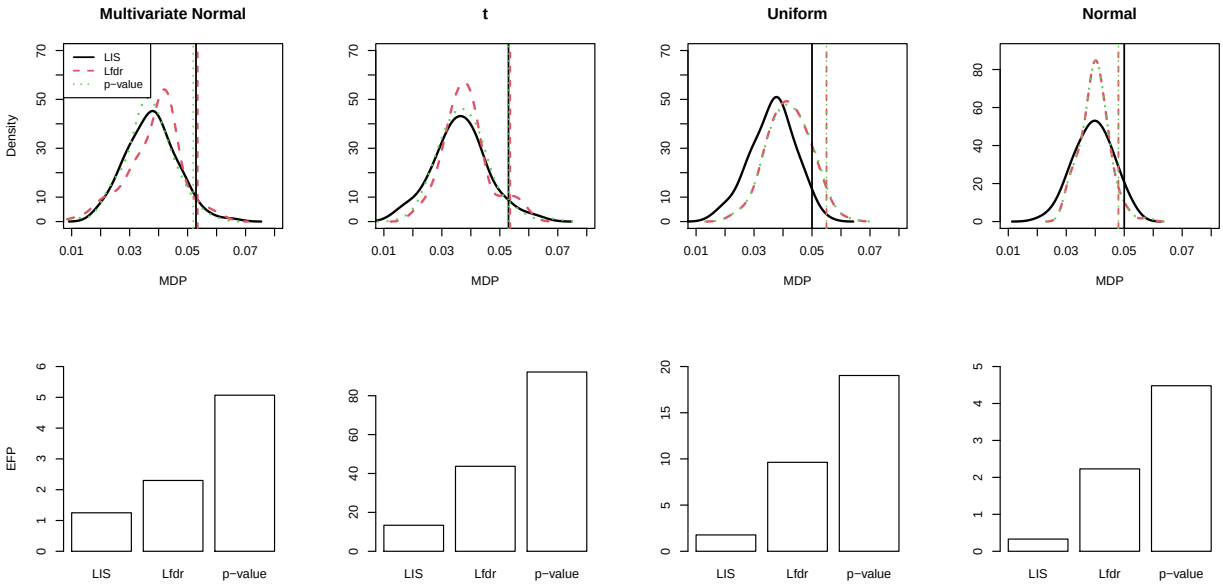


Figure 5: Sensitivity analysis of the proposed data-driven diagnostic procedure for MDX-control against: HMM (first column), marginal normality (second and third columns) and degenerate OC distribution (fourth column). The first row shows the density curves of MDP and the second row shows the EFP levels. The vertical lines in the plots of the first row denote the $100(1-\alpha)$th percentiles of MDP of the MDX-control procedure using different thresholding approaches.

shown in the first column of Figure 5, from which we can see that MDX is still controlled at the nominal level by our procedure, and thresholding LIS can obtain the smallest EFP among the three thresholding choices.

Second, we study the sensitivity of our method to the marginal normality assumption. To this end, the following two distributions are considered:

- $t$ distribution: $x_i|\theta_i \sim t(4) + 2\theta_i$,
- Uniform distribution: $x_i|\theta_i \sim \theta_i U(0,4) + (1-\theta_i)U(-2,2)$,

where $\boldsymbol{\theta}$ is generated based on the HMM model in the strong correlation case. The simulation results are shown in the second and third columns of Figure 5. It can be seen as expected that our procedure for MDX-control is quite robust against non-normal distributions, mainly because it is based on the sample mean of $n$ OC observations.

Third, we study the sensitivity of our method to the assumption of degenerate OC

distribution. To this end, the OC observations are generated from the following model:

$$x_i|\mu_i \sim N(\mu_i, 1),$$

where $\mu_i|\theta_i \sim \theta_i U(1.5, 2.5)+(1-\theta_i)\delta_0(\mu_i)$, $\delta_0(\cdot)$ is the Dirac delta function, and $\boldsymbol{\theta}$ is generated from the HMM model in the strong correlation case. The simulation results are shown in the fourth column of Figure 5, from which we can have similar conclusions to those in other cases in this example that our procedure for MDX-control is robust against non-degenerate OC distribution. Therefore, this example shows that our proposed diagnostic procedure for MDX-control is quite robust against model misspecifications, which is appealing for practical use.

**Diagnostic Performance Under A False Alarm.** In SPC, control charts could trigger false OC signals in cases when the related process is actually IC. In this part, we study the diagnostic performance of the proposed data-driven procedure for MDX-control under a false alarm. In such a case, the indices MDP and EFP become meaningless since there are no OC data streams with $\theta_i = 1$. Instead, we focus on the expected number of positives (EP), defined to be $E(\sum_{i=1}^{m} \delta_i)$. In the example of Figure 2, set $m = 1,000$ and $n = 10$, and consider the strong correlation case. The EP values of the data-driven procedure for MDX-control with various combinations of $\alpha$ and $\tau$ based on 10,000 simulations are summarized in Table 3. From the table, it can be seen that the EP values are all very small ($< 17$), compared to the value of $m = 1,000$, in all scenarios considered, which implies that only a few data streams would be diagnosed mistakenly by the proposed method as OC. Therefore, it can be concluded that the proposed diagnostic procedure is quite robust under a false alarm.

Table 3: EP values of the data-driven procedure for MDX-control under a false alarm. Their standard errors are included in parentheses.

| | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|
| $\tau = 0.01$ | 16.6(1.98) | 11.5(1.25) | 9.4(1.03) |
| $\tau = 0.05$ | 13.1(1.46) | 8.5(0.94) | 7.2(0.77) |
| $\tau = 0.1$ | 10.2(1.17) | 6.7(0.70) | 4.8(0.46) |

# 5 Application to a Semiconductor Manufacturing Dataset

The modern semiconductor manufacturing process (SMP) is one of the most complicated processes nowadays. The increased complexity of high-tech devices gives a tremendous challenge for proper quality control. SMP typically involves a series of complicated steps, and the key variables in the whole process are monitored constantly based on data streams collected from hundreds of sensors. As a motivating example, we analyze a real SMP dataset available from the UCI Machine Learning Repository. It contains 1,463 conforming samples and 104 nonconforming samples, and each sample is a high-dimensional observation of 590 variables. These variables are ordered by the manufacturing procedural steps that they belong to. For demonstration purposes, the 1,463 conforming samples are regarded as the IC dataset and the 104 nonconforming samples are regarded as the OC dataset for fault diagnosis in this section.

When the SMP changes from IC to OC, the mean values of certain data streams may shift abruptly. The timely and accurate fault diagnosis of the OC status is thus critically important to identify the OC data streams and repair the manufacturing process accordingly. High-dimensional fault diagnosis involves simultaneous testing of a large number of data streams, where good reliability and sensitivity are among the top considerations. To this end, we demonstrate how the proposed data-driven diagnostic procedure can be applied to this SMP example for identifying the OC data streams.

To analyse the data properly, pre-processing of the data is often needed. First, the constant data streams and the extremely discrete data streams can be deleted from the analysis, resulting in $m = 453$ remaining streams for further analysis. Also, because there are a small number of missing values in the data, the mean imputation approach is used to substitute each missing value with the mean value of the corresponding data stream. Such imputation would not change the sample mean for each data stream. The Shapiro-Wilks test for normality indicates that many data streams are not normally distributed. Thus, the transformation $\Phi^{-1}(\widehat{F}_k(X_{kt}))$, for $k = 1, ..., m$, is implemented to each data stream, where $\widehat{F}_k$ is the empirical c.d.f. of the $k$th data stream estimated from the IC dataset, $X_{kt}$ is the observed value of the $k$th data stream at the time point $t$, and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal c.d.f.. After this transformation, each data stream would have a

distribution that is close to normal, but the joint distribution of all $m$ data streams may not be normal.

The observed data are assumed to be a normal mixture of two hidden patterns, with one for the IC data streams with the normal distribution $N(\mu_{IC}, \sigma^2)$ and the other for the OC data streams with the distribution $N(\mu_{OC}, \sigma^2)$. The hidden states of the data streams $\boldsymbol{\theta}$ are assumed to form a Markov chain. The HMM parameters are estimated by using the EM algorithm. The estimation results with $n = 20$ are summarized in Table 4, where $\widehat{a}_{11} = 0.89$ indicates that a positive correlation exists among the hidden statuses, and that the OC streams tend to appear in clusters. It should be noted that if $\boldsymbol{x}$ follows an HMM-based mixture distribution, then the estimation results in Table 4 by using the EM algorithm should be close to the true underlying distribution. Motivated by this intuition, we performed the chi-square goodness-of-fit test to see whether sample frequencies follow the estimated mixture distribution well. Specifically, sample data are divided into intervals, and the numbers of sample data in individual intervals are compared with the expected numbers of sample data under the HMM model by constructing the chi-square test statistic. The resulting $p$-value is 0.216 in this case, which implies that the observed data can be described properly by the HMM model.

The proposed data-driven LIS-based diagnostic procedure for MDX-control is then applied to the SMP data. After setting $\alpha = \tau = 0.05$, the diagnostic results are displayed in Figure 6, where the estimated LIS values are plotted. The threshold value $c^{**} = 0.7556$ is obtained by using the proposed Monte Carlo method. It can be seen that a total of 301 data streams are identified as OC, and most of them appear in clusters, which indicates that the hidden states of the data streams are correlated.

Table 4: The estimated mixture model for the SMP data.

| IC distribution | OC distribution | Transition matrix $\boldsymbol{A}$ |
|---|---|---|
| $N(-0.05, 0.42^2)$ | $N(0.25, 0.42^2)$ | $\begin{pmatrix} 0.94 & 0.06 \\ 0.11 & 0.89 \end{pmatrix}$ |

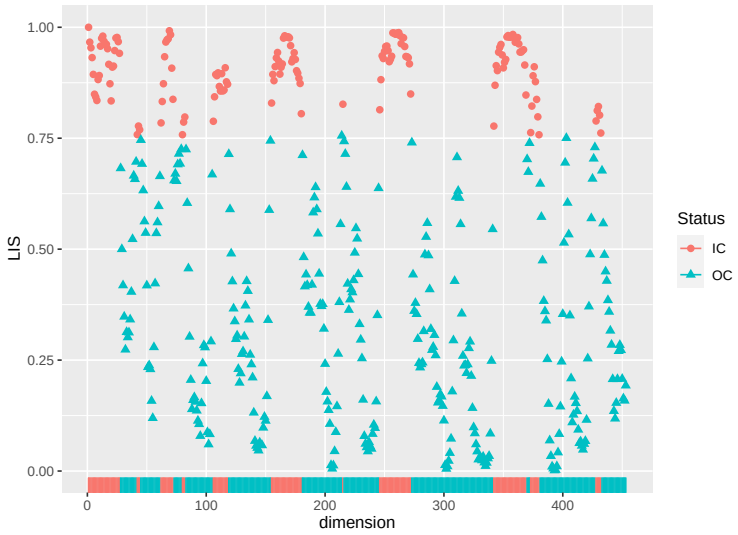


Figure 6: The diagnostic results of the SMP data by using the data-driven procedure for MDX-control.

# 6 Concluding Remarks

In this paper, we have focused on the reliable fault diagnosis problem for high-dimensional and mutually correlated data streams, which is formulated into a large-scale multiple testing framework for controlling MDX under the HMM dependence structure. Both the oracle and data-driven diagnostic procedures for MDX-control are discussed based on LIS thresholding. A Monte-Carlo method is proposed for estimating MDX, and then the optimal threshold value can be found numerically by using the bisection search algorithm. Based on extensive simulation results and a real-data analysis, we can see that the proposed procedures for MDX-control are reliable and effective for fault diagnosis of high-dimensional data streams.

Several issues still need to be studied in the future research. First, the proposed diagnostic procedures for MDX-control are designed mainly for detecting abnormal mean shifts. After certain modifications, they should be able to diagnose covariance shifts as well, which has not be discussed in the paper yet. Second, the HMM assumption might be strong for certain applications. Much future research is needed to develop reliable diagnostic procedures under more flexible data dependence structures. Third, in many real-world applications, it should be important to accommodate both spatial and temporal data correlations in the context of fault diagnosis, which will be one of the main research directions in the future. To this end, our proposed multiple-testing based method could be combined with the existing

data decorrelation procedures (e.g., Qiu et al., 2020; Li and Qiu, 2020; Xue and Qiu, 2020; Qiu and Xie, 2021). But, this is beyond the scope of the current paper, and will be studied elsewhere. Finally, SPC of network and image data becomes more and more important in real applications (e.g., Dong et al., 2020; Ebrahimi et al., 2020; Feng and Qiu, 2018; Menafoglio et al., 2018; Qiu 2020; Wang and Xie, 2021). The problem of fault diagnosis for network and image data is challenging, which also needs be studied in the future.

# Supplementary Materials

**Supplementary File:** The Supplementary File contains additional simulation results.

**Code and Data:** The Code and Data contain codes for performing MDX-control and the real dataset.

# Acknowledgments

The authors thank the editor, the associate editor and three referees for many constructive comments and suggestions, which greatly improved the quality of the paper. This research was supported in part by National Science Foundation [DMS-1914639], National Natural Science Foundation of China [12071144; 71931004; 11871324; 11801210; 11771145], National Science Foundation of Shanghai [19ZR1414400], China Postdoctoral Science Foundation [2020M671064], National Bureau of Statistics of China [2020LD03], and the Fundamental Research Funds for the Central Universities.

# References

Apley D.W. and Tsung, F. (2002). The autoregressive $T^2$ chart for monitoring univariate auto-correlated processes, *Journal of Quality Technology*, 34, 80–96.

Baum, L., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occuring in

the statistical analysis of probabilistic functions of Markov Chains, *Annals of Mathematical Statistics*, 41, 164–171.

Bickel, P., Ritov, Y. and Rydén, T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models, *Annals of Statistics*, 26, 1614–1635.

Capizzi, G. and Masarotto, G. (2011). A least angle regression control chart for multidimensional data, *Technometrics*, 53(3), 285–296.

Dong, H., Chen, N. and Wang, K. (2020). Modeling and change detection for count-weighted multilayer networks, *Technometrics*, 62(2), 184–195.

Durrett, R. (2005). *Probability: Theory and Examples*, 3rd edition, Belmont: Duxbury.

Ebrahimi, S., Reisi Gahrooei, M., Manakad, S. and Paynabar, K. (2020). Monitoring sparse and attributed networks with online hurdle models. *IISE Transactions*, 1–31.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis, *Journal of the American Statistical Association,* 99, 96–104.

Ephraim, Y. and Merhav, N. (2002). Hidden Markov processes, *IEEE Transactions on Information Theory*, 48, 1518–1569.

Feng, L. and Qiu, P. (2018). Difference detection between two images for image monitoring, *Technometrics*, 60, 345–359.

Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101, 1408–1417.

Kim, J. , Jeong, M. K., Elsayed, E. A., Al-Khalifa, K. N. and Hamouda, A. M. S. (2016). An adaptive step-down procedure for fault variable identification, *International Journal of Production Research*, 54(11), 3187–3200.

Leroux, B. (1992). Maximum-likelihood estimation for hidden Markov models, *Stochastic Processes & Their Applications*, 40, 127–143.

Li, J., Jin, J. and Shi, J. (2008). Causation-based T2 decomposition for multivariate process monitoring and diagnosis. *Journal of Quality Technology*, 40(1), 46–58.

Li, W., Pu, X., Tsung, F. and Xiang, D. (2017). A robust self-starting spatial rank multivariate EWMA chart based on forward variable selection, *Computers & Industrial Engineering*, 103, 116–130.

Li, W. and Qiu, P. (2020). A general charting scheme for monitoring serially correlated data with short-memory dependence and nonparametric distributions, *IISE Transactions*, 52(1), 61–74.

Li, W., Xiang, D., Tsung, F. and Pu, X. (2020a). A diagnostic procedure for high-dimensional data streams via missed discovery rate control, *Technometrics*, 62, 84–100.

Li, W., Zhang, C., Tsung, F. and Mei, Y. (2020b). Nonparametric monitoring of multivariate data via KNN learning, *International Journal of Production Research*, https://urldefense.proofpoint.com/v2/url?u=https-3A__doi.org_10.1080_00207543.2020.1812750&d=DwIGAg&c=sJ6xIWYx-zLMB3EPkvcnVg&r=PGwSI8-Ygwtyl5QNS8xJ1KbfJiOWy-BLUj_xRBxS62Y&m=lNgVM0KuCAi79vwnMTugvwrTF_IQjUjnD7YOHF6XHCc&s=0RwW6PGWIRA83TrIQyTa1Z-xL2YtCU8_6cEK8lJD9bw&e=.

Liu, K., Mei, Y. and Shi, J. (2015). An adaptive sampling strategy for online high-dimensional process monitoring, *Technometrics*, 57(3), 305–319.

Mason, R. L., Tracy, N. D. and Young, J. C. (1995). Decomposition of T2 for multivariate control chart interpretation, *Journal of Quality Technology*, 27(2), 109–119.

Mason, R. L., Tracy, N. D. and Young, J. C. (1997). A practical approach for interpreting multivariate T2 control chart signals, *Journal of Quality Technology*, 29(4), 396–406.

Menafoglio, A., Grasso, M., Secchi, P. and Colosimo, B. M. (2018). Profile monitoring of probability density functions via simplicial functional PCA with application to image data, *Technometrics*, 60, 497-510.

Pacifico, M. P., Genovese, C., Verdinelli, I. and Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association*, 99, 1002–1014.

Qiu, P. (2014). *Introduction to Statistical Process Control*, Boca Raton, FL: Chapman & Hall/CRC.

Qiu, P. (2020). Big data? statistical process control can help!, *The American Statistician*, 74, 329–344.

Qiu, P., Li, W. and Li, J. (2020). A new process control chart for monitoring short-range serially correlated data. *Technometrics*, 62(1), 71–83.

Qiu, P. and Xie, X. (2021). Transparent sequential learning for statistical process control of serially correlated data. *Technometrics*, https://urldefense.proofpoint.com/v2/url?u=https-3A__doi.org_10.1080_00401706.2021.1929493&d=DwIGAg&c=sJ6xIWYx-zLMB3EPkvcnVg&r=PGwSI8-Ygwtyl5QNS8xJ1KbfJiOWy-BLUj_xRBxS62Y&m=lNgVM0KuCAi79vwnMTugvwrTF_IQjUjnD7YOHF6XHCc&s=cjbTyByzIIySf1mGkYGMBBtRh5yXJ0iPkq4UQ7eNAwE&e=.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77, 257–286.

Samuel, T.R. and Pignatiello, Jr., J.J. (2001). Estimation of the change point of a normal process mean in SPC applications, *Journal of Quality Technology*, 33, 82–95.

Sullivan, J. H., Stoumbos, Z. G., Mason, R. L. and Young, J. C. (2007). Step-down analysis for changes in the covariance matrix and other parameters, *Journal of Quality Technology*, 39(1), 66–84.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control, *Journal of the American Statistical Association*, 102, 901–912.

Sun, W. and Cai, T. T. (2009). Large-scale multiple testing under dependence, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71, 393–424.

Wang, K. and Jiang, W. (2009). High-dimensional process monitoring and fault isolation via variable selection, *Journal of Quality Technology*, 41(3), 247–258.

Wang, J. and Xie, M. (2021). Modeling and monitoring unweighted networks with directed interactions. *IISE Transactions*, 53(1), 116–130.

Xiang, D., Li, W., Tsung, F., Pu, X. and Kang, Y. (2021). Fault classification for high-dimensional data streams: A directional diagnostic framework based on multiple hypothesis testing. *Naval Research Logistics*, https://urldefense.proofpoint.com/v2/url?

u=https-3A__doi.org_10.1002_nav.22008&d=DwIGAg&c=sJ6xIWYx-zLMB3EPkvcnVg&r=PGwSI8-Ygwtyl5QNS8xJ1KbfJiOWy-BLUj_xRBxS62Y&m=lNgVM0KuCAi79vwnMTugvwrTF_IQjUjnD7YOHF6XHCc&s=p0I2BUYgwn5hegPl5QhAfJHeyzPeytoSiKZZQu0Sm4k&e=.

Xue, L. and Qiu, P. (2020). A nonparametric CUSUM chart for monitoring multivariate serially correlated processes. *Journal of Quality Technology*, https://urldefense.proofpoint.com/v2/url?u=https-3A__doi.org_10.1080_00224065.2021.1903820&d=DwIGAg&c=sJ6xIWYx-zLMB3EPkvcnVg&r=PGwSI8-Ygwtyl5QNS8xJ1KbfJiOWy-BLUj_xRBxS62Y&m=lNgVM0KuCAi79vwnMTugvwrTF_IQjUjnD7YOHF6XHCc&s=TGcc-x3yJo31l9sp-HPJ2oXZemsNbzTKNB9Fmm1K8CA&e=.

Yan, H., Paynabar, K. and Shi, J. (2018). Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition, *Technometrics*, 60(2), 181–197.

Zamba, K. D. and Hawkins, D. M. (2006). A multivariate change-point for statistical process control, *Technometrics*, 48, 539–549.

Zhang, C., Chen, N. and Wu, J. (2020). Spatial rank-based high-dimensional monitoring through random projection, *Journal of Quality Technology*, 52(2), 111-127.

Zhu, Y. and Jiang, W. (2009). An adaptive T2 chart for multivariate process monitoring and diagnosis, *IIE Transactions*, 41(11), 1007–1018.

Zou, C., Jiang, W. and Tsung, F. (2011). A lasso-based diagnostic framework for multivariate statistical process control, *Technometrics*, 53(3), 297–309.

Zou, C. and Qiu, P. (2009). Multivariate statistical process control using lasso, *Journal of the American Statistical Association*, 104, 1586–1596.

Zou, C., Wang, Z., Jiang, W. and Zi, X. (2015). An efficient online monitoring method for high-dimensional data streams, *Technometrics*, 57(3), 374–387.