Received: 13 June 2020 | Revised: 24 March 2021 | Accepted: 22 May 2021 | Published on: 11 June 2021

DOI: 10.1002/nav.22008

RESEARCH ARTICLE

WILEY

Fault classification for high-dimensional data streams: A directional diagnostic framework based on multiple hypothesis testing

Dongdong Xiang[1] | **Wendong Li**[2] | **Fugee Tsung**[3] | **Xiaolong Pu**[1] | **Yicheng Kang**[4]

[1]KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China
[2]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China
[3]Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Kowloon, Hong Kong
[4]Department of Mathematical Science, Bentley University, Waltham, USA

Correspondence
Wendong Li, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China.
Email: wendongli01@gmail.com

Funding information
National Bureau of Statistics of China, Grant/Award Number: 2020LD03. China Postdoctoral Science Foundation, Grant/Award Number: 2020M671064. Fundamental Research Funds for the Central Universities. National Natural Science Foundation of China, Grant/Award Numbers: 12071144, 71931004, 11871324, 11771145, 71931006, 11801210. National Science Foundation of Shanghai, Grant/Award Number: 19ZR1414400. Research Grants Council, University Grants Committee, Grant/Award Numbers: 16201718, 16216119.

Abstract

In various modern statistical process control applications that involve high-dimensional data streams (HDDS), accurate fault diagnosis of out-of-control (OC) streams is becoming crucial. The existing diagnostic approaches either focus on moderate-dimensional processes or are unable to determine the shift direction accurately, especially when the signal-to-noise ratio is low. In this paper, we conduct a bold trial and consider the *fault classification* problem of the mean vector of HDDS where determining the shift direction of the OC streams is important to perform customized repairs. To this end, under the basic assumptions that the in-control data streams are normal with mean 0 and variance 1, and that the high-dimensional observations after the alarm are solely OC, the problem is formulated into a three-classification multiple testing framework, and an efficient data-driven diagnostic procedure is developed to minimize the expected number of false positives and to control the missed discovery rate at given level. The procedure is statistically optimal and computationally efficient, and improves the diagnostic effectiveness by considering directional information, which provides insights to guide further decisions. Both theoretical and numerical results reveal the superiority of the new method.

KEYWORDS

data-driven, directional isolation, high-dimensional fault diagnosis, multiple testing, statistical process control

1 | INTRODUCTION

With the rapid development of automatic in-process measurement and data capture techniques, high-dimensional data streams (HDDS) consisting of high-dimensional and sequential continuous observations have become highly common in industrial applications, posing great challenges to the field of conventional multivariate statistical process control (MSPC). MSPC basically includes two tasks. The first one is to make real-time decision whether the process has changed from in-control (IC) to out-of-control (OC), and is often referred to as online monitoring. The second one, post-signal fault diagnosis, is to isolate the abnormal data streams responsible for the change. When the process has undergone abnormal changes, accurate fault diagnosis is crucial to help practitioners eliminate the root causes of the OC state. While the online monitoring problem of HDDS has attracted considerable attention recently (Li, Zhang, et al., 2020; Liu et al., 2015; Mei, 2010; Xian et al., 2018; Yan et al., 2018; Zhang et al., 2020; Zou et al., 2015), the fault diagnosis problem of HDDS is still an open field that needs to be studied systematically. In this paper, we focus on the fault diagnosis problem for the mean vector of HDDS.

Fault diagnosis of HDDS is closely related to applying conventional MSPC diagnostic procedures to m variables (assuming the data dimension is m), a topic that has been fully investigated. The earliest studies attempted to capture the relationship between different process parameters

by interpreting and decomposing Hotelling's T^2-type statistics (Li et al., 2008; Mason et al., 1997), based on which various step-down methods were also proposed (Sullivan et al., 2007; Zhu & Jiang, 2009). Recently, by using variable selection (VS) techniques, several MSPC schemes mainly designed for online monitoring have been proposed (Capizzi & Masarotto, 2011; Li et al., 2017; Wang & Jiang, 2009; Zou & Qiu, 2009), which can be applied for a rough fault diagnosis. Zou et al. (2011) proposed a unified multivariate diagnostic framework that combines BIC and adaptive LASSO, which has been shown to have better performance than other conventional methods in various applications. The aforementioned multivariate diagnostic approaches are intuitively sound, but may be suboptimal when applied to HDDS. The major challenge is that, with complex underlying models and high dimensionality (much larger than the number of OC observations), the curse of dimensionality problem arises. Besides, these methods are computationally intensive because they involve large-scale matrix computation. To handle HDDS effectively, recently Zhang et al. (2020) proposed a diagnostic framework based on the square-root LASSO algorithm, yet their work is mainly concentrated in online monitoring. Li, Xiang, et al. (2020) suggested a procedure to isolate a subset with the minimum amount of IC information and almost all of the OC information.

All the diagnostic procedures discussed above are designed for nondirectional fault diagnosis, that is, they can only isolate the shifted OC data streams, but they are not necessarily able to determine the shift direction (positive or negative) of each OC data stream. However, in many applications, the causes underlying the shifts in different directions may differ, and engineers must adopt customized measures based on the shift direction. This is the case for many data sets arising from manufacturing industries. A concrete motivating example is the modern manufacturing process of semiconductor integrated circuits and devices, which involves a series of complicated steps. The key variables in the manufacturing process are monitored persistently based on high-dimensional data streams collected from numerous automatic sensors (May & Spanos, 2006). When the process becomes OC, it is of great importance to identify the data streams responsible for the OC state, and carry out repairs as soon as possible. Engineers must determine the shift direction of the OC data streams, as the repair method to be adopted will depend on the shift direction. Performing customized repairs based on the shift direction will increase product quality and decrease the manufacturing cost.

As a result, conducting directional fault diagnosis, or *fault classification*, to determine the shift direction systematically is highly desirable for HDDS. Unfortunately, existing methods often overlook directional information at all, and can only determine the shift direction from the sign of the observed value, which is too inaccurate for fault classification, and consequently leads to poor diagnosis power when applied to HDDS, especially when signal-to-noise ratio (SNR) is low. In such circumstances, the precious OC information is often drowned out by the IC data streams and noise, posing great challenges to further decisions. Currently there is no satisfactory directional diagnostic approach for HDDS. The fault classification problem of HDDS still remains an open field ripe for exploration.

In this paper, we aim to fill this research gap by proposing an effective fault classification framework for HDDS to assist in the isolation of data streams that are responsible for the abnormal changes. To focus on the diagnostic phase, similar to Zou et al. (2011) and Li, Xiang, et al. (2020), we assume without loss of generality that the IC streams are normally distributed with mean 0 and variance 1, and that the high-dimensional observations after the alarm are solely abnormal. We formulate the fault classification problem of HDDS on the basis of a three-classification large-scale multiple testing framework. Based on the framework, a novel definition of the missed discovery rate (MDR) is proposed. A directional diagnostic procedure that minimizes the expected number of false positives (EFP) and controls the MDR at given level is then developed, into which the directional information of the OC data streams are organically integrated. We establish theoretically its validity and optimality for fault classification of HDDS. The proposed diagnostic procedure and the corresponding theoretical results are also extended to the cases when between-stream correlation exists. The numerical performance of the proposed diagnostic procedure is investigated via extensive simulation studies and a real life example. The numerical results imply that the proposed procedure outperforms its rivals in various scenarios.

The remainder of this paper is organized as follows. The formulation of the fault classification problem of HDDS is introduced in Section 2. The oracle and data-driven procedures for fault classification and their validity and optimality are developed in Section 3. Simulation studies are given in Section 4. Section 5 applies the proposal to a real-data example. Several concluding remarks are given in Section 6. Technical details are given in the Online Appendix.

2 | FAULT CLASSIFICATION PROBLEM OF HDDS

In this section, we describe the fault classification problem of HDDS in detail and formulate it into a multiple testing framework. It is well known in the statistical process control (SPC) literature that after a change point, the underlying process will change from IC to OC. When the process is IC, at each time point t, there is an observation of dimension m collected, denoted by $\boldsymbol{X}_t$. Without loss of generality, it is assumed that $\boldsymbol{X}_t$s are i.i.d., and that the data streams are standardized beforehand with mean 0 and variance 1. When the process becomes OC, certain online monitoring scheme will give an alarm signal. After the signal, assume that $n \geq 1$

observations $\boldsymbol{X}_1^{OC}, \ldots, \boldsymbol{X}_n^{OC}$ are collected with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^T$. In such a situation, a small set of components of $\boldsymbol{\mu}$ are non-zero due to the OC pattern. We further assume that the distribution of $\boldsymbol{X}_i^{OC}$s is:

$$\begin{aligned} \boldsymbol{X}_j^{OC}|\boldsymbol{\mu}, \boldsymbol{\theta} &\sim N_m(\boldsymbol{\mu}, \Sigma)\mu_i|\theta_i \sim (1 - |\theta_i|)\delta_0(\mu_i) \\ &\quad + I(\theta_i = 1)h_1(\mu_i) + I(\theta_i = -1)h_2(\mu_i) \\ \theta_i &\overset{i.i.d}{\sim} \text{Multinoulli}(p_0, p_1, p_{-1}), \\ &\sum_{k=0,1,-1} p_k = 1,\ i = 1, \ldots, m,\ j = 1, \ldots, n, \end{aligned} \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$ denotes the state (IC, positive OC, or negative OC with probability p_0, p_1, or p_{-1}, respectively) of each data stream, $\delta_0(\cdot)$ is the Dirac delta function, $h_1(\mu), \mu > 0$ and $h_2(\mu), \mu < 0$ are the probability density functions of μ_i given $\theta_i = 1$ and $\theta_i = -1$, respectively. Note that variations of (1) have been widely used in the field of high-dimensional analysis (Cai & Sun, 2009; Efron, 2004; Mei, 2010; Zou et al., 2015). Based on $\boldsymbol{X}_1^{OC}, \ldots, \boldsymbol{X}_n^{OC}$, we utilize the sample mean $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^n \boldsymbol{X}_j^{OC}$ to implement fault classification, which is a simple yet effective way to enhance SNR. The notation $\overline{\boldsymbol{X}}$ is simplified as $\boldsymbol{X} = (X_1, \ldots, X_m)^T$ in this paper for convenience. Also note that the assumption of normality in model (1) may be violated in real applications. Nevertheless, for the most common non-normal distributions in real applications, X_is would be asymptotically normal as n increases. In such cases, procedures derived under the normality assumption should still work well, as will be discussed numerically in later sections.

In essence, the objective of the fault classification problem is to isolate the non-zero values of $\boldsymbol{\mu}$, determine their shift directions, and repair the corresponding data streams based on their shift directions. Since θ_i takes the value 1 if $\mu_i > 0$, -1 if $\mu_i < 0$, and 0 otherwise, the objective is equivalent to a three-classification multiple testing problem:

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

The solution to (2) can be represented by a set of decision functions, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^T \in \{0,1,-1\}^m$, where $\delta_i = 1, -1$, or 0 if we claim that the ith data stream is positive OC, negative OC, or IC respectively. Based on $\boldsymbol{\delta}$, the results of applying a three-classification multiple testing method are presented in Table 1. Appropriate indices should then be defined based on this table to evaluate the diagnostic performance of $\boldsymbol{\delta}$.

To solve the fault classification problem of HDDS, we need to ensure that most of the OC data streams are discovered and the right direction identified. In such a context, the missed discovery rate is defined as $\text{MDR}(\boldsymbol{\delta}) = E(\frac{N_{10}+N_{12}+N_{20}+N_{21}}{m_1+m_2})$. A closely related concept is the marginal MDR, which is defined by

$$\text{mMDR}(\boldsymbol{\delta}) = \frac{E(N_{10} + N_{12} + N_{20} + N_{21})}{E(m_1 + m_2)}.$$

TABLE 1 Classification of tested hypothesis

	$\delta_i = 0$	$\delta_i = 1$	$\delta_i = -1$	Total
IC ($\theta_i = 0$)	N_{00}	N_{01}	N_{02}	m_0
OC ($\theta_i = 1$)	N_{10}	N_{11}	N_{12}	m_1
OC ($\theta_i = -1$)	N_{20}	N_{21}	N_{22}	m_2
Total	R_0	R_1	R_2	m

In this paper, we utilize the mMDR to construct the diagnostic framework, mainly for considerations to obtain theoretical optimality results (see Section 3 for more details). It should be noted that such marginal idea has been widely applied in the multiple testing literature (Cai et al., 2019; Sun & Cai, 2007). For convenience, the mMDR is abbreviated as MDR in the rest of the paper.

We emphasize that the definition of MDR implies that when an OC stream with $\theta_i = 1$ is discovered but deemed as $\delta_i = -1$, it should still be considered as a missed discovery. If the MDR equals α, then $(1 - \alpha) \times 100$ percent of the OC data streams are discovered and the right direction identified. While controlling the MDR, setting a reasonable target is also crucial for the evaluation of $\boldsymbol{\delta}$. In high-dimensional SPC applications, since discovering too many IC data streams is undesirable, the target is defined as the expected number of false positives

$$\text{EFP}(\boldsymbol{\delta}) = E(N_{01} + N_{02}).$$

Based on the above definitions, the fault classification problem of HDDS now becomes a constrained optimization problem

$$\min_{\boldsymbol{\delta}} \text{EFP}(\boldsymbol{\delta}) \text{ subject to } \text{MDR}(\boldsymbol{\delta}) \le \alpha. \tag{3}$$

Before ending this section, we would like to remark on the fault classification problem of HDDS. Our formulation (1)–(3) is based on the compound decision-making framework that involves solving m decision problems simultaneously. In the literature of multiple testing, many procedures have been proposed based on the compound decision-making framework to minimize the false non-discovery rate (FNR) and to control the false discovery rate (FDR) or its variants at given level (Holte et al., 2016; Sun & Cai, 2007). However, in SPC, these procedures may not fit into the fault classification problem of HDDS. The main reason is that the FDR being controlled only guarantees that the proportion of the IC data streams that are mis-specified as OC is controlled. It is still possible that many OC data streams would be missed, which runs counter to the objective of fault classification in industrial applications. Noticing that no existing diagnostic procedure is designed for the fault classification problem of HDDS, our formulation of minimizing the EFP while controlling the MDR can be regarded as a perfect fusion of the fault classification problem and the directional compound decision-making framework.

3 | MULTIPLE TESTING-BASED DIRECTIONAL DIAGNOSTIC FRAMEWORK FOR HDDS

In this section, we set out to the following tasks: (i) propose the oracle procedure for the fault classification of HDDS and prove its validity and optimality for MDR-control, where oracle means that the distribution of $\boldsymbol{X}$ is completely known; and (ii) to address practical needs, propose the data-driven procedure and show its asymptotic validity and optimality. Following the convention in high-dimensional SPC analysis, we first assume that the data streams are independent of each other to derive our method in Sections 3.1–3.2, that is, $\Sigma = I_{m\times m}$, where $I_{m\times m}$ is the m-dimensional identity matrix. The extension to the correlated case and the robustness of the proposed diagnostic procedure will be discussed in Section 3.3 and also numerically in Section 4.

3.1 | Oracle diagnostic procedure

Let $g(X_i) = p_0 g(X_i|\theta_i = 0) + p_1 g(X_i|\theta_i = 1) + p_{-1} g(X_i|\theta_i = -1)$ be the marginal probability density function (p.d.f.) of X_i, where $g(X_i|\theta_i = k)$ is the marginal p.d.f. of X_i given $\theta_i = k$. Define

$$H_k(X_i) = P(\theta_i = k|X_i) = g(X_i|\theta_i = k)P(\theta_i = k)/g(X_i). \quad (4)$$

Then we can easily check that under model (1), problem (3) is equivalent to minimizing

$$\begin{aligned}\mathrm{EFP}(\boldsymbol{\delta}) &= E\left[\sum_{i=1}^{m}|\delta_i|(1-|\theta_i|)\right]\\ &= E_X\left[\sum_{i=1}^{m}\sum_{k=1,-1} I(\delta_i = k)H_0(X_i)\right]\end{aligned}$$

subject to

$$E_X\left\{\sum_{i=1}^{m}\sum_{k=1,-1}[1 - I(\delta_i = k) - \alpha]\,H_k(X_i)\right\} \le 0,$$

where E_X is the expectation taken over X_i. To solve this constrained optimization problem, the following penalized objective function is considered:

$$\begin{aligned}L_X(\lambda, \boldsymbol{\delta}) = \sum_{i=1}^{m}\sum_{k=1,-1}\{&I(\delta_i = k)H_0(X_i)\\ &+ \lambda\,[1 - I(\delta_i = k) - \alpha]\,H_k(X_i)\}, \quad (5)\end{aligned}$$

where λ is a penalty parameter. Given $\lambda > 0$, we can easily verify that $L_X(\lambda, \boldsymbol{\delta})$ is minimized by

$$\delta_i^{\lambda} = \begin{cases} k, & \text{if } \Lambda_i(X_i) = \frac{H_{max}(X_i)}{H_0(X_i)} \ge 1/\lambda \text{ and}\\ & H_{max}(X_i) = H_k(X_i), \text{ for } k = \pm 1\\ 0, & \text{Otherwise,}\end{cases} \quad (6)$$

where $H_{max}(X_i) = \max\{H_1(X_i), H_{-1}(X_i)\}$. Therefore, $E_X(L_X(\lambda, \delta))$ is also minimized. We show in Theorem 1 below the validity and optimality of the decision rule (6) for fault classification of HDDS.

Theorem 1 *Under model* (1), *consider the oracle statistics* $H_k(X_i)$ *and the penalized objective function* $L_X(\lambda, \boldsymbol{\delta})$. *Denote by* $Q_{OR}(\lambda)$ *the level of decision rule* $\boldsymbol{\delta}^{\lambda} = \{\delta_1^{\lambda}, \ldots, \delta_m^{\lambda}\}$, *where* δ_i^{λ} *is defined in* (6). *Define*

$$\lambda^* = \inf\{\lambda : Q_{OR}(\lambda) \le \alpha\}.$$

Then for any $\alpha \ge \lim_{\lambda\to\infty} Q_{OR}(\lambda)$, *we have*

i. λ^* *exists uniquely, and* $\mathrm{MDR}(\boldsymbol{\delta}^{\lambda^*}) = \alpha$.
ii. *For any* $\boldsymbol{\delta}$ *satisfying* $\mathrm{MDR}(\boldsymbol{\delta}) \le \alpha$, $\mathrm{EFP}(\delta^{\lambda^*}) \le \mathrm{EFP}(\boldsymbol{\delta})$.

It is worth noting that the condition $\alpha \ge \lim_{\lambda\to\infty} Q_{OR}(\lambda)$ is indispensable in the fault classification problem of HDDS. The reason is that this condition is essentially related to our definition of the MDR, in which a data stream with $\theta_i = 1(-1)$ determined as $\delta_i = -1(1)$ is flagged as a miss discovery, whereas in the nondirectional fault diagnosis problem, only a data stream with $\theta_i \ne 0$ determined as $\delta_i = 0$ is flagged as a miss discovery. Obviously, the difficulty of the former is more significant than that of the latter, especially when the OC signal is weak. In other words, when the shift size is close to 0, it is generally difficult to classify faults and determine their shift directions, making missed discoveries inevitable to some degree. To better understand the nature of this issue, we rewrite $Q_{OR}(\lambda)$ as follows:

$$\begin{aligned}Q_{OR}(\lambda) = \mathrm{MDR}(\boldsymbol{\delta}^{\lambda}) &= 1 - \frac{E_X\left\{\sum_{i=1}^{m}\sum_{k=1,-1} I(\delta_i^{\lambda} = k)H_k(X_i)\right\}}{E_X\left[\sum_{i=1}^{m}\sum_{k=1,-1} H_k(X_i)\right]}\\ &= 1 - \frac{E_X\left\{\sum_{i=1}^{m} H_{max}(X_i)\mathrm{I}\left[\Lambda_i(X_i) \ge 1/\lambda\right]\right\}}{E_X\left[\sum_{i=1}^{m}\sum_{k=1,-1} H_k(X_i)\right]}.\end{aligned}$$

It can be easily observed that, for any λ

$$\mathrm{MDR}(\boldsymbol{\delta}^{\lambda}) \ge 1 - \frac{E_X\left[\sum_{i=1}^{m} H_{max}(X_i)\right]}{E_X\left[\sum_{i=1}^{m}\sum_{k=1,-1} H_k(X_i)\right]} \triangleq LB > 0,$$

where $LB = \lim_{\lambda\to\infty} Q_{OR}(\lambda)$ is the lower bound of the MDR of $\boldsymbol{\delta}^{\lambda}$. Intuitively, if the shift size is close to 0 (i.e., low SNR), then the probabilities of $\theta_i = 1$ and -1 given X_i would be similar, and $H_1(X_i)$ and $H_{-1}(X_i)$ would have similar values. In such a situation, H_{max} would be much smaller than $H_1 + H_{-1}$, and LB would be large, indicating that if we choose an α that is too small, then it would be impossible to maintain the MDR at α. As the shift size increases (i.e., higher SNR), H_1 would deviate widely from H_{-1} and LB would be close to 0, which gives us more choices of α for valid fault classification. Therefore, with a very weak OC signal and a small α, we must collect more OC observations than usual to control the MDR. Further numerical examples are given in Section 4 for illustration.

The optimal threshold λ^* is often unavailable since the MDR is unknown in practice, making the oracle decision rule $\boldsymbol{\delta}^{\lambda^*}$ inoperable. Fortunately, under mild conditions, we can

derive an asymptotically equivalent diagnostic procedure as follows. First, we can estimate the MDR by the following moment estimator:

$$\widehat{\mathrm{MDR}}(\boldsymbol{\delta}^{\lambda}) = 1 - \frac{\sum_{i=1}^{m} H_{max}(X_i)\mathrm{I}(\Lambda_i(X_i) \geq 1/\lambda)}{\sum_{i=1}^{m}\sum_{k=1,-1} H_k(X_i)} \triangleq 1 - B.$$

Thus, to ensure that $\widehat{\mathrm{MDR}}(\boldsymbol{\delta}^{\lambda}) \leq \alpha$, we need to verify that $B \geq 1-\alpha$. Denote by Γ the oracle statistics $\Lambda_i(X_i)$s in increasing order: $\Gamma = \left\{\Lambda_{(1)}, \ldots, \Lambda_{(m)}\right\}$. By substituting Γ into the indicator functions in B, $\boldsymbol{\delta}^{\lambda^*}$ can be approximated as follows.

Procedure 1 (Stepwise Oracle Procedure for MDR-control, SOM)**.** Rank $\Lambda_i(X_i)$ in increasing order to obtain $\Gamma = \left\{\Lambda_{(1)}, \ldots, \Lambda_{(m)}\right\}$. Let

$$\gamma = \max\left\{ j : \frac{\sum_{i=j}^{m} H_{max}^{(i)}}{\sum_{i=1}^{m}\sum_{k=1,-1} H_k(X_i)} \geq 1-\alpha \right\}.$$

Now we reject $H_{(i)}^0$ *for* $i = \gamma, \ldots, m$, where $H_{(i)}^0$ is the null hypothesis and $H_{max}^{(i)}$ has a one-to-one correspondence with $\Lambda_{(i)}$. Then the shift direction is determined by deciding whether $H_{max}^{(i)}$ equals H_1 or H_{-1}.

It can be checked that

$$\Lambda^*(X_i) \triangleq \frac{1}{1+\Lambda_i(X_i)} = \frac{p_0 g(X_i|\theta_i=0)}{p_0 g(X_i|\theta_i=0) + \max\{p_1 g(X_i|\theta_i=1), p_{-1} g(X_i|\theta_i=-1)\}}$$

is monotonically decreasing in $\Lambda_i(X_i)$, and the equation can be regarded as a novel generalization of the local FDR (Efron, 2004)

$$\mathrm{Lfdr} = \frac{p_0 g(X_i|\theta_i=0)}{p_0 g(X_i|\theta_i=0) + p_1 g(X_i|\theta_i=1) + p_{-1} g(X_i|\theta_i=-1)}.$$

Although the Lfdr has been widely applied in the literature to interpret results for individual cases, it is nondirectional and thus suboptimal for fault classification in the sense that the positive and negative shifts are not considered separately, as can be observed from the denominator of the Lfdr. In some ways, $\Lambda^*(X_i)$ (or $\Lambda_i(X_i)$ equivalently) is a directional version of the Lfdr which utilizes the directional information and the idea of maximum likelihood ratio for directional decision. Therefore, $\Lambda_i(X_i)$ is more powerful and appropriate than the Lfdr in the fault classification problem of HDDS.

Theorem 2 below shows that the SOM procedure is asymptotically valid and optimal for MDR-control.

Theorem 2 *Under model* (1)*, consider the oracle statistics* $H_k(X_i)$ *and* $\Lambda_i(X_i)$*. Let* $\gamma = \max\left\{ j : \frac{\sum_{i=j}^{m} H_{max}^{(i)}}{\sum_{i=1}^{m}\sum_{k=1,-1} H_k(X_i)} \geq 1-\alpha \right\}$ *and* $\boldsymbol{\delta}^S = \left\{\delta_i^S, i = 1, \ldots, m\right\}$ *with*

$$\delta_i^S = \begin{cases} k, & \text{if } \Lambda_i(X_i) \geq \Lambda_{(\gamma)} \text{ and } H_k(X_i) \geq H_{k'}(X_i) \text{ for} \\ & k, k' = \pm 1, k \neq k'. \\ 0, & \text{Otherwise.} \end{cases}$$

For any $\alpha \geq lim_{\lambda\to\infty} Q_{OR}(\lambda)$*, we have*

i. $\mathrm{MDR}(\boldsymbol{\delta}^S) = \alpha + o(1)$.
ii. $\mathrm{EFP}(\boldsymbol{\delta}^S)/\mathrm{EFP}(\boldsymbol{\delta}^{\lambda^*}) = 1 + o(1)$.

3.2 | Data-driven diagnostic procedure

When developing SOM, it is assumed that the underlying model (1) is completely known, and thus SOM cannot be applied for practical purposes. To solve this problem, we now propose a data-driven procedure for fault classification of HDDS. In order to construct a data-driven procedure, a natural idea is to find appropriate estimates of $H_k(X_i), k = 0,1,-1$ to give plug-in estimators $\widehat{H}_{max}(X_i)$, $\widehat{H}_0(X_i)$ and $\widehat{\Lambda}_i(X_i)$. In what follows, we discuss the estimation of the quantities in $H_k(X_i)$s based on the OC sample mean $\boldsymbol{X}$ and give practical guidelines.

First, estimate $\widehat{g}(X_i)$ for the denominator of $H_k(X_i)$, the marginal density function $g(X_i)$, can be obtained with conventional kernel-based methods (Silverman, 1986). Now the numerators of $H_k(X_i)$s need to be estimated. To estimate the numerator of $H_0(X_i)$, it suffices to estimate the null proportion p_0, since $g(X_i|\theta_i = 0)$ is the density function of standard normal. A consistent estimate $\widehat{p}_0$ for p_0 can be obtained immediately by using the estimation method proposed by Jin and Cai (2007). Finally, for the estimation of the numerators of $H_1(X_i)$ and $H_{-1}(X_i)$, $p_k g(X_i|\theta_i = k), k = \pm 1$ must be estimated. In general, we need to estimate p_k and $g(X_i|\theta_i = k)$ separately. Fortunately, we find after derivations that $p_k g(X_i|\theta_i = k), k = \pm 1$ can be written as

$$p_1 g(X_i|\theta_i = 1) = (1-p_0)\int_0^{+\infty} g(X_i|\theta_i \neq 0, \mu)h(\mu)\mathrm{d}\mu$$

and

$$p_{-1} g(X_i|\theta_i = -1) = (1-p_0)\int_{-\infty}^{0} g(X_i|\theta_i \neq 0, \mu)h(\mu)\mathrm{d}\mu,$$

where

$$h(\mu) = \frac{p_1}{1-p_0}h_1(\mu) + \frac{p_{-1}}{1-p_0}h_2(\mu)$$

is the density function of μ_i given $\theta_i \neq 0$. Therefore, in order to estimate $p_k g(X_i|\theta_i = k), k = \pm 1$, we only need to estimate $h(\mu)$, and $p_k g(X_i|\theta_i = k)$ can be estimated by numerical approximation of the integrals, since $g(X_i|\theta_i \neq 0, \mu)$ is the density function of $N(\mu, 1)$. Such approach is easy to implement and also of high accuracy. For estimating $h(\mu)$, the deconvoluting kernel estimator (Sun & McLain, 2012) can be used:

$$\widehat{h}(\mu) = \frac{1}{2\pi(1-\widehat{p}_0)}\int_{-\infty}^{\infty} e^{-it\mu}\left[\widehat{\Psi}(t)/\Psi_{\varepsilon}(t) - \widehat{p}_0\right]\Psi_K(\tau t)dt, \tag{7}$$

where $\widehat{\Psi}(t)$ is the empirical characteristic function of $\boldsymbol{X}$, Ψ_{ε} and Ψ_K are the characteristic functions of the error distribution and a kernel $K(t)$, respectively, and τ is the bandwidth parameter. In practice, estimator (7) is often transformed into $\max\left\{0, \widehat{h}(\mu)\right\}$. In (7), it is important to choose the kernel function $K(t)$ properly. A related discussion has been provided

by Delaigle and Hall (2006), who study kernel selection in deconvolution problems systematically. In this paper, the sinc kernel $K(t) = (\pi t)^{-1} \sin t$ with $\Psi_K(t) = I(|t| \leq 1)$ is chosen for the normal distribution. To estimate the bandwidth parameter τ, we utilize the bandwidth selection method proposed by Delaigle and Gijbels (2004), in which the bootstrap-based approximated MISE of h is minimized. By using this bandwidth selection method, the proposed data-driven procedure roughly achieves the optimal performance as demonstrated by our empirical results.

Now, given $\widehat{H}_{max}(X_i)$, $\widehat{H}_0(X_i)$ and $\widehat{\Lambda}_i(X_i)$, the data-driven diagnostic procedure is given as follows. Its asymptotic validity and optimality for MDR-control are presented in Theorem 3 in the Online Appendix.

Procedure 2 (Stepwise Data-driven Procedure for MDR-control, SDM)**.** Rank $\widehat{\Lambda}_i(X_i)$ in increasing order to obtain $\left\{\widehat{\Lambda}_{(1)}, \ldots, \widehat{\Lambda}_{(m)}\right\}$. Let

$$\gamma = \max\left\{j : \frac{\sum_{i=j}^{m} \widehat{H}_{max}^{(i)}}{\sum_{i=1}^{m} \sum_{k=1,-1} \widehat{H}_k(X_i)} \geq 1-\alpha\right\}. \quad (8)$$

Now we reject $H_{(i)}^0$ for $i = \gamma, \ldots, m$, where $H_{(i)}^0$ is the null hypothesis and $\widehat{H}_{max}^{(i)}$ has a one-to-one correspondence with $\widehat{\Lambda}_{(i)}$. Then the shift direction is determined by deciding whether $\widehat{H}_{max}^{(i)}$ equals $\widehat{H}_1$ *or* $\widehat{H}_{-1}$.

3.3 | Extension to dependent data streams

The diagnostic procedures proposed in Sections 3.1–3.2 are based on the independence assumption in the sense that the data streams are assumed to be independent of each other, that is, $\boldsymbol{\Sigma} = I_{m\times m}$ in model (1). However, in real applications, this assumption can be invalid, and between-stream correlation exists. In this subsection, we further extend the proposed diagnostic procedures to allow the data streams to be correlated. In such a situation, the joint oracle statistics turn out to be

$$H_{ki}(\boldsymbol{X}) = P(\theta_i = k|\boldsymbol{X}) = g(\boldsymbol{X}|\theta_i = k)P(\theta_i = k)/g(\boldsymbol{X}),$$
$$k = 0,1,-1,$$

where $g(\boldsymbol{X})$ is the joint p.d.f. of X and $g(\boldsymbol{X}|\theta_i = k)$ is the joint p.d.f. given $\theta_i = k$. Based on $H_{ki}(\boldsymbol{X})$, we can define $\Lambda_i(\boldsymbol{X})$ and then derive the joint oracle and data-driven procedures for MDR-control similar to Procedures 1 and 2. Their theoretical validity and optimality for MDR-control can also be established.

However, when the data streams are correlated, calculating the joint p.d.f.s, $g(\boldsymbol{X}|\theta_i = k)$ and $g(\boldsymbol{X})$, is computationally too expensive to afford. Specifically, the computational complexity of calculating the oracle statistics $H_{ki}(\boldsymbol{X})$ is $O(m2^m)$. Fortunately, in practice, information on the correlation structure is often known in advance. For instance, short-range correlation structure has been widely applied to characterize correlation between data streams in high-dimensional cases (Xiang et al., 2019). In the literature, an effective and efficient solution for handling certain weak between-stream correlation structures is to neglect it, which can significantly speed up computation without sacrificing performance much. For example, Qiu et al. (2010) proposed a Phase II profile monitoring scheme that considers the heteroscedasticity of observations and ignores within-profile correlation. Similar ideas can also be found in the longitudinal data analysis literature (Lin & Carroll, 2000) and the multiple testing literature (Xie et al., 2011). In our case, we can similarly prove that the proposed marginal data-driven procedure SDM in Section 3.2 is still valid and optimal asymptotically for MDR-control, which is shown in Theorem 4 in the Online Appendix. More detailed numerical investigation of the diagnostic performance under correlation are given in Section 4.2. Section 6 also discusses the generalization of SDM under more complicated correlation structures.

4 | SIMULATION STUDIES

In this part, we assess numerically the diagnostic performance of the proposed stepwise oracle procedure for MDR-control, SOM, and its data-driven version, SDM. There are basically no comparable methods designed for fault classification of HDDS in the literature. It has been shown by Li, Xiang, et al. (2020) that their proposed diagnostic oracle/data-driven procedures MOW/MDW outperform other conventional MSPC diagnostic procedures. Therefore, we modify the MOW and MDW procedures so that they can be applied for the fault classification of HDDS. To be more specific, after the diagnostic step in which the MOW and MDW procedures isolate the OC data streams, the shift directions of these OC data streams are simply determined from the signs of the observations.

We set $\alpha = 0.1$ for all the considered methods, which is reasonable in practice. The simulation results for other α are similar, and thus are omitted to save space. After $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ are generated, the actual MDR and EFP values of the diagnostic procedures are obtained from 1000 replications. Then we repeat the whole process, from fixing $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ to obtaining the MDR and EFP values, 100 times, after which we can obtain the mean values of the MDR and EFP. Specifically, in Section 4.1, we investigate the performance of the considered procedures when model (1) is true and the data streams are independent of each other. In Section 4.2, we consider cases when between-stream correlation exists. Section 4.3 studies the impact of m. Finally, we investigate the robustness of the procedures when the normality assumption is violated in Section 4.4.

4.1 | I.I.D. normal cases

In this part, we investigate the effectiveness and robustness of the proposed diagnostic procedures under the normal mixture

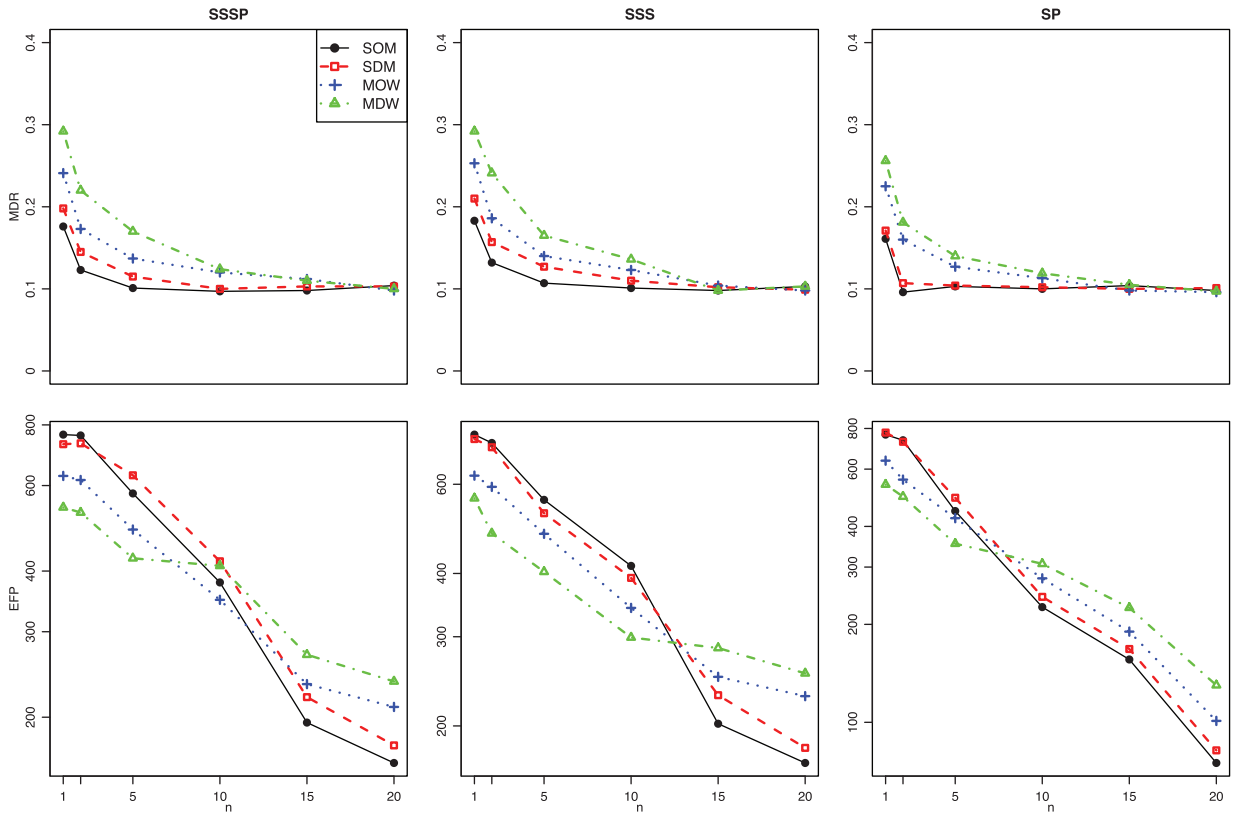

FIGURE 1 Diagnostic results of the i.i.d. normal cases for SOM (●), SDM (□), MOW (+), and MDW (▲). The MDR and EFP levels are respectively shown in the top and bottom rows. The scenarios are at the top of the columns

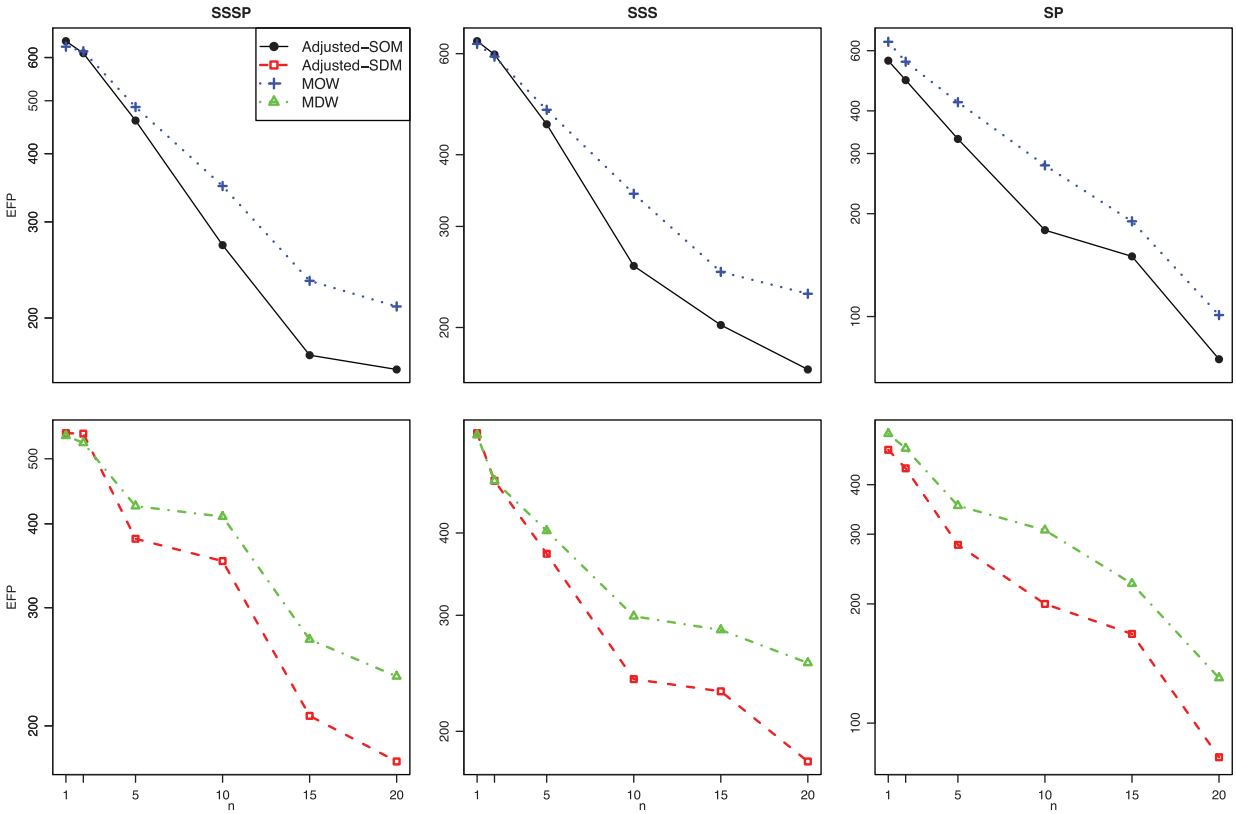

FIGURE 2 EFP levels of the four diagnostic procedures after MDR adjustment

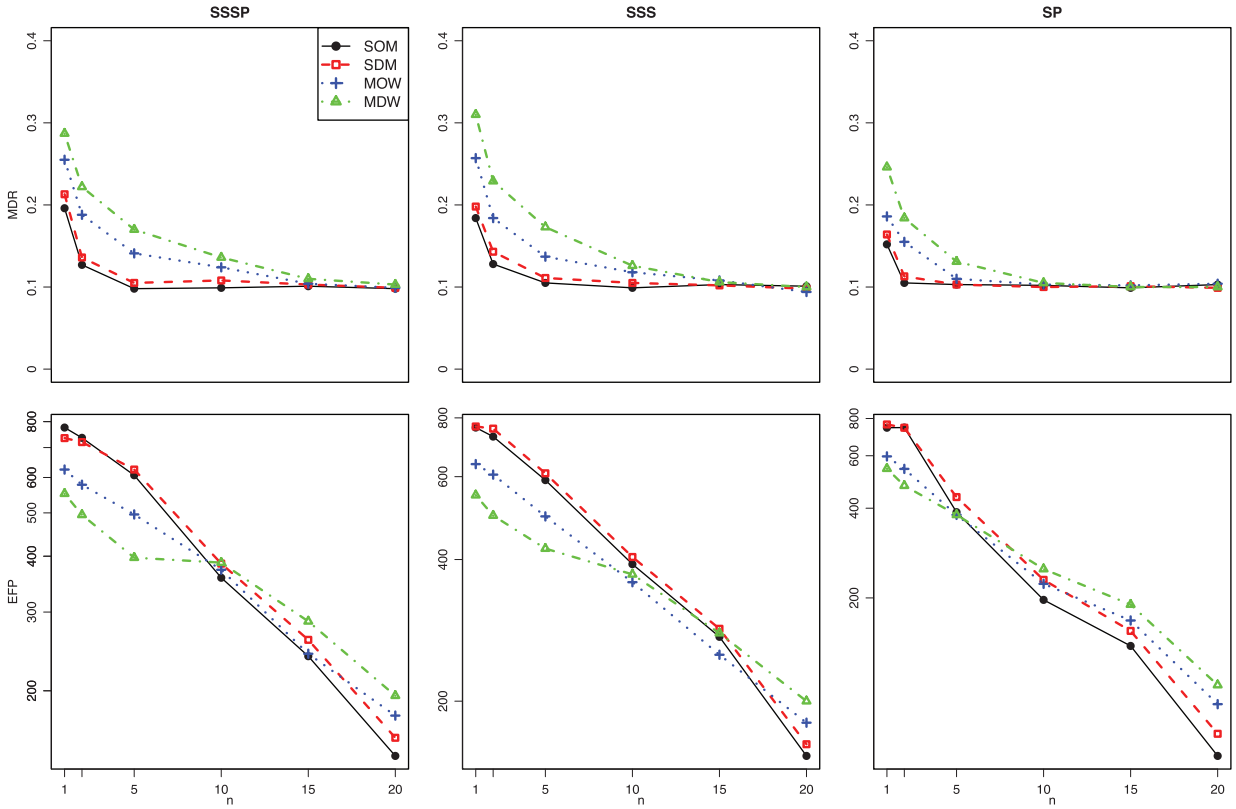

FIGURE 3 Diagnostic results of Case 1 for SOM (●), SDM (□), MOW (+), and MDW (▲). The MDR and EFP levels are respectively shown in the top and bottom rows. The scenarios are at the top of the columns

model (1) and when the data streams are independent. We fix $m = 1000$ and $p_0 = 0.75$, and consider the following three scenarios:

- Symmetric OC shift sizes and proportions (SSSP): $h_1(\mu_i) = h_2(-\mu_i) = Gam(2,0.05,0.5)$ and $p_1 = p_{-1} = 0.125$, where $Gam(A1, B1, C1)$ denotes the gamma distribution with shape parameter $A1$, location parameter $B1$ and scale parameter $C1$.
- Symmetric OC shift sizes (SSS): $h_1(\mu_i) = h_2(-\mu_i) = Gam(2,0.05,0.5)$, $p_1 = 0.1$ and $p_{-1} = 0.15$.
- Symmetric OC proportions (SP): $h_1(\mu_i) = Gam(3,0.05,0.5), h_2(-\mu_i) = Gam(2,0.05,0.5)$ and $p_1 = p_{-1} = 0.125$.

In Scenario SSSP, the OC shift sizes and proportions both have a symmetric structure, while only one of them is symmetric in Scenarios SSS and SP. The simulation results are displayed in Figure 1. For demonstration purposes, we plot the MDR and EFP as functions of the number of OC observations, n, where n changes among 1, 2, 5, 10, 15, and 20. Note that the scale on the y-axis of the EFP plots is in natural logarithm, to better demonstrate the difference among procedures.

From the plots, we can observe that when n is very small, none of the four procedures can control the MDR at 0.1, but the proposed SOM and SDM procedures have much smaller MDR values than MOW and MDW do. This also validates the discussion in Section 3.1 that we cannot control the MDR with a low SNR and a small α. The MDR values of SOM and SDM converge to 0.1 rapidly as n increases, while neither MOW nor MDW are valid until n is very large (e.g., $n = 20$). After the convergence, SOM and SDM are both valid, and the EFP values of SDM are slightly larger than these of SOM, implying that the performance of SOM is asymptotically attained by SDM. By contrast, in cases when the MDR levels of MOW and MDW are very close to 0.1, their EFP levels are significantly larger than those of SOM and SDM, respectively. This shows the sub-optimality of the MOW and MDW procedures for the fault classification problem of HDDS.

In order to see more thoroughly the advantages of the proposed diagnostic procedures, we artificially adjust the actual MDR levels of SOM and SDM to be the same as those of MOW and MDW, respectively, and the EFP levels after the adjustment are displayed in Figure 2. Note that the oracle and data-driven procedures are compared separately. From the plots, we can observe that SOM and SDM significantly outperform MOW and MDW, respectively. The differences become more substantial as n gets larger. Therefore, after taking into account its computational advantage, we believe that the SDM procedure provides an effective tool for fault classification of HDDS.

4.2 | Impact of between-stream correlation

In this part, we consider the cases when between-stream correlation exists. Specifically, we choose the following two

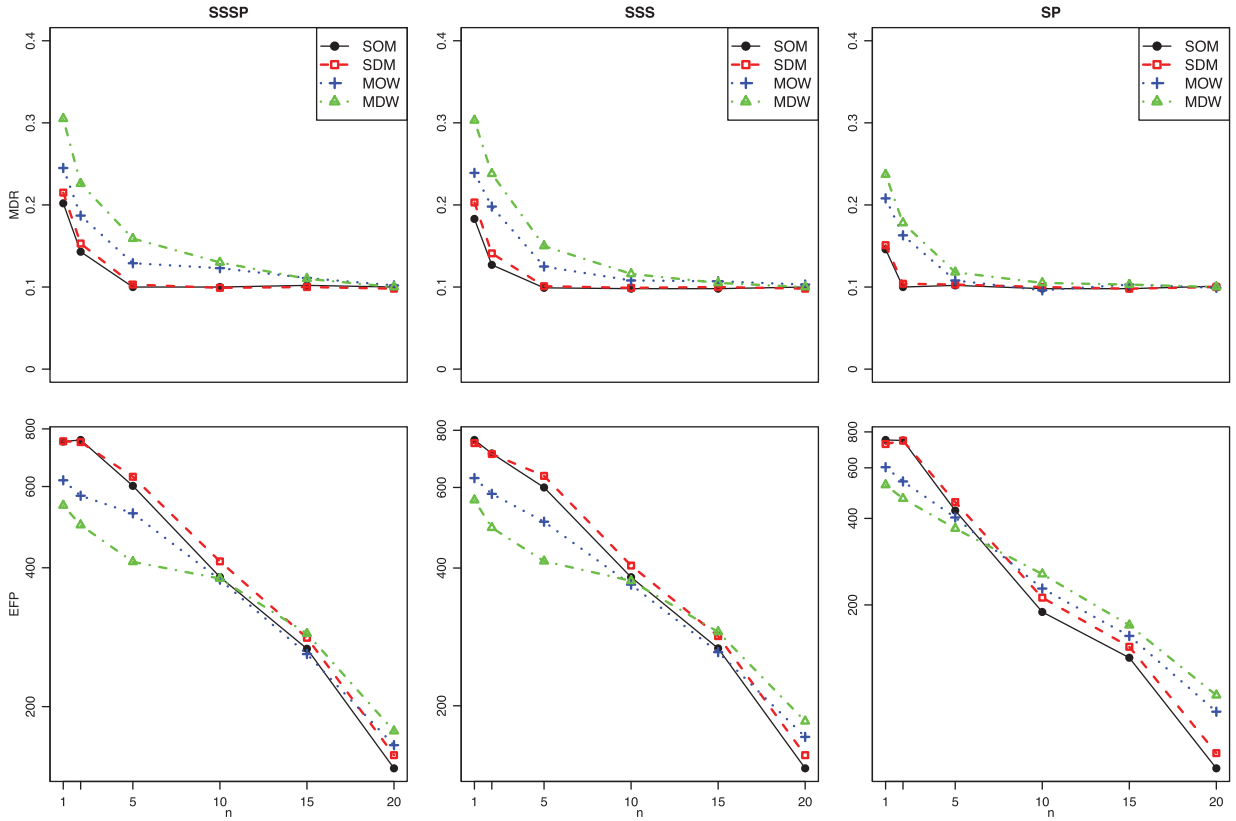

FIGURE 4 Diagnostic results of Case 2 for SOM (●), SDM (□), MOW (+), and MDW (▲). The MDR and EFP levels are respectively shown in the top and bottom rows. The scenarios are at the top of the columns

representative correlation structures:

Case 1 $\mathbf{\Sigma}$ is block-diagonal. The block size is equal to $b = 10$ with the diagonal elements of 1 and the off-diagonal elements of 0.3 in each block.

Case 2 $\mathbf{\Sigma} = (\sigma_{ij})_{m\times m} = \rho^{|i-j|}$ with $\rho = 0.5$.

Case 1 is a short-range correlation structure, and the magnitude of correlation in Case 2 decays as the position moves away from the diagonal. The other settings are identical to those in Section 4.1. Then the OC observations can be generated under model (1). The diagnostic results of Cases 1 and 2 are summarized and displayed in Figures 3 and 4, respectively. From the plots, similar conclusions can be drawn: the SOM procedure performs the best among the other three procedures, indicating its validity and optimality for MDR-control. Also, the performance of the SOM procedure is asymptotically attained by the SDM procedure, and SOM and SDM outperform MOW and MDW, respectively.

Next, we study in more detail the impact of correlation on the diagnostic performance of the proposed procedures SOM and SDM. To this end, we consider the block-diagonal covariance matrix in Case 1 with b changing among 0, 10, 100 and 500, and the covariance matrix in Case 2 with ρ changing among 0, 0.1, 0.5 and 0.9. To facilitate presentation, only the SP scenario is displayed, and the results of the other two scenarios are similar based on our empirical results. The simulation results are displayed in Figure 5. From the plots, we could draw the following conclusions. First, SOM has strong robustness against various degrees of correlation in the sense that the MDR and EFP values are very similar under different values of b or ρ. This should not be surprising, since the marginal distributional information is completely known for SOM regardless of the correlation structure. Second, in cases when weak correlation exists (i.e., $b \leq 10$ or $\rho \leq 0.5$), it can be observed that SDM performs satisfactorily, as the MDR values are controlled at 0.1, and the EFP values are similar to these when the data streams are independent (i.e., $b = 0$ or $\rho = 0$). These findings are consistent with the theoretical results established in Section 3. Third, in cases when the correlation is strong (i.e., $b \geq 100$ or $\rho > 0.5$), it can be observed that SDM is still valid for directional MDR-control, as the MDR level converges to 0.1 as n gets larger. However, the EFP values are much larger than those of the independent and weak correlated cases, indicating that the effectiveness of SDM is seriously affected by the strong correlation. The main reason may be that it is extremely difficult to precisely estimate the marginal distributional information under such strong correlation.

Finally, it can be concluded that the proposed diagnostic procedures SOM and SDM are still very effective under certain weak or short-range correlation structures. When the correlation is extremely strong or long-range, the effectiveness of SOM and SDM may be seriously affected. In such

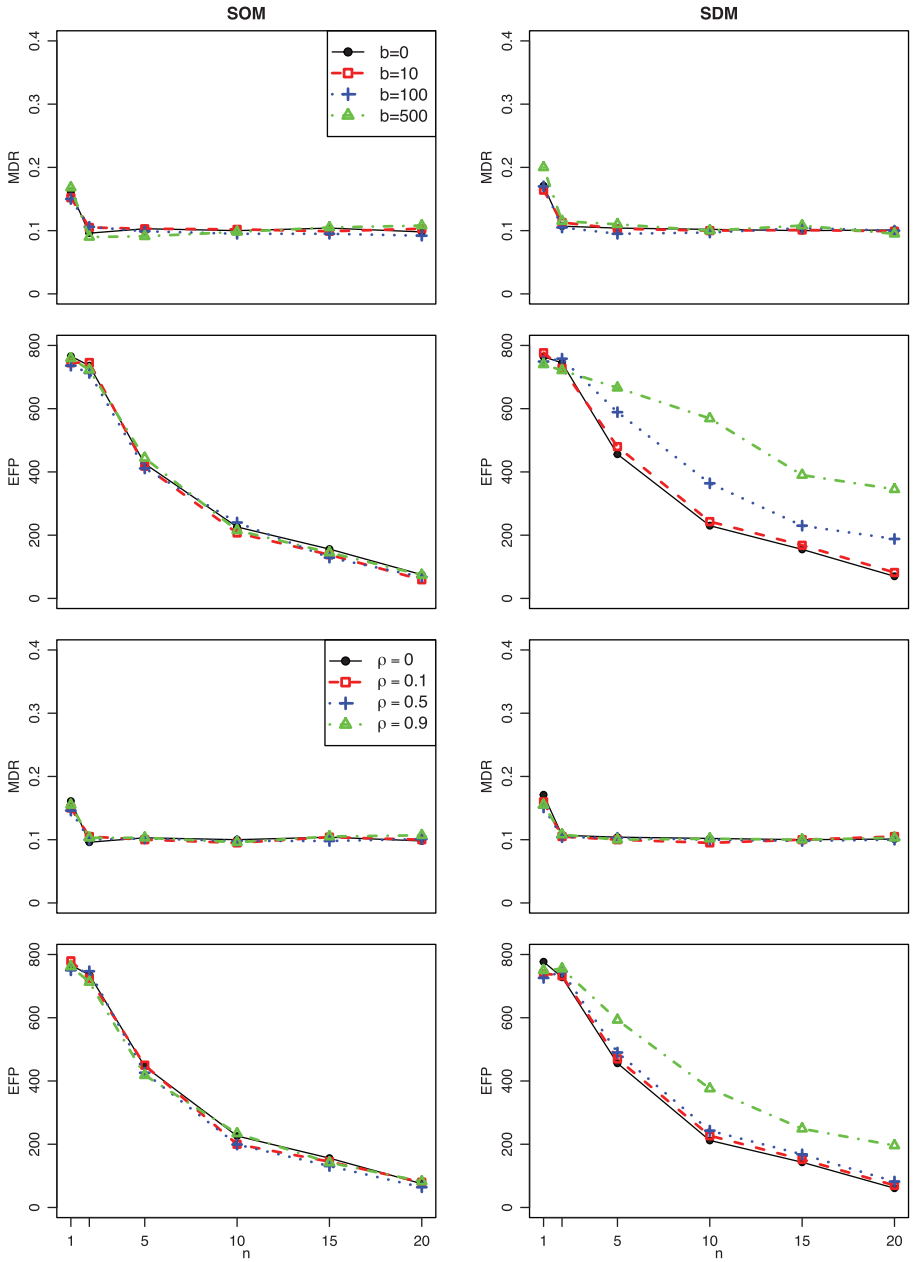

FIGURE 5 Diagnostic performance of SOM (the first column) and SDM (the second column) under various degrees of correlation. The first two rows are the MDR and EFP values with b changing among 0, 10, 100, and 500, and the second two rows are the MDR and EFP values with ρ changing among 0, 0.1, 0.5, and 0.9

cases, more general and effective diagnostic procedures are desirable to take into consideration the correlation properly to facilitate decision making. See also Section 6 for more discussions.

4.3 | Impact of dimensionality

Based on the theoretical results in Section 3, we can conclude that the performance of the proposed SOM and SDM procedures would depend on the dimensionality, m through the $o(1)$ terms in the theorems in Section 3. Intuitively, provided that m is sufficiently large, our procedures can work satisfactorily. In this subsection, such potential impact is studied numerically. We let m change among 500, 1000, and 2000, and consider the short-range correlation structure (Case 1 in Section 4.2) and the SP scenario. The diagnostic performance of SOM and SDM are displayed in Figure 6. For demonstration, all the EFP values are divided by the dimensionality m.

From the plots, the following conclusions can be drawn: (i) The SOM and SDM procedures are both valid for MDR-control for all choices of m as their MDR levels converge to α when n increases, (ii) the diagnostic performance of both SOM and SDM would improve when m increases as the EFP level tends to be smaller. The improvement of SDM is more significant, mainly because that it is hard to estimate the parameters accurately when m is small, (iii) the diagnostic performance stabilizes when $m \geq 1000$, and the improvement can almost be ignored. This numerical example shows that the

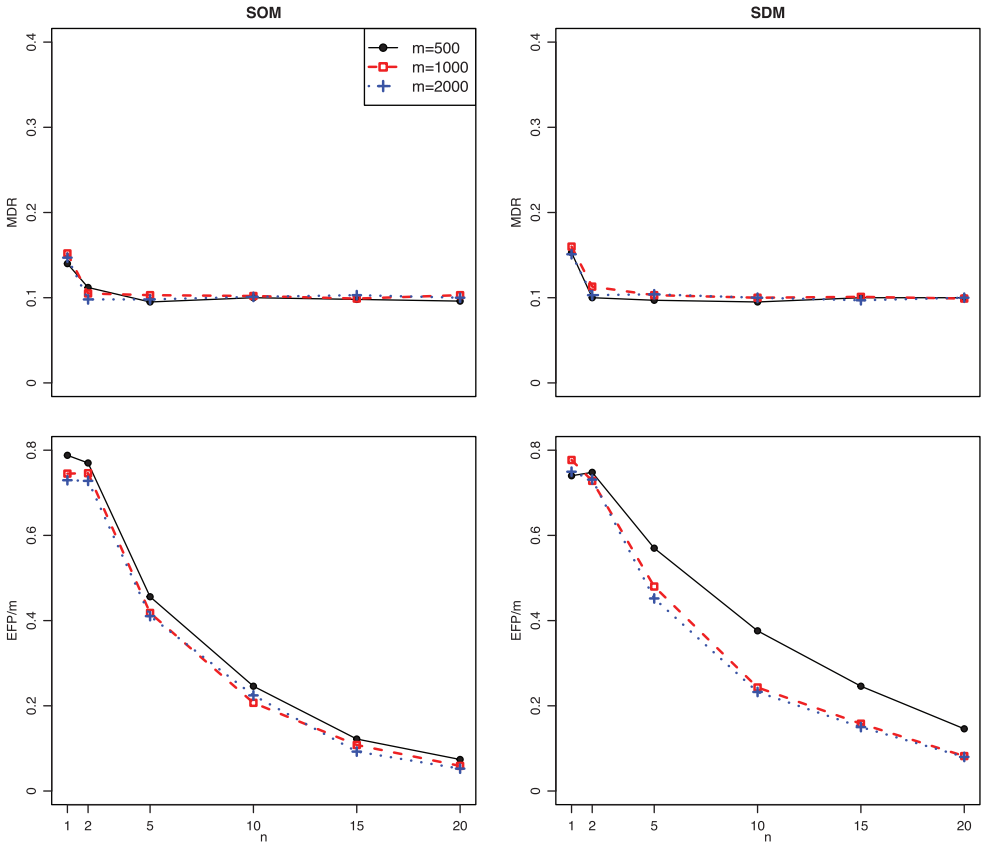

FIGURE 6 Diagnostic performance of SOM (the first column) and SDM (the second column) with m changing among 500, 1000, and 2000. The MDR and EFP levels are respectively shown in the top and bottom rows

proposed SOM and SDM procedures perform stably as the dimensionality is as large as 1000.

4.4 | Impact of non-normality

The foregoing numerical examples are all based on model (1) with the assumption of normality. In this subsection, we study the performance of the procedures without the normality assumption. As discussed in Section 2, provided that n is not too small, the distribution of X_is should be asymptotically normal. In such cases, procedures derived under the normality assumption should still work well. To verify this, we consider the following two non-normal distributions: t_3 and $Gam(2,0,1)$. For simplicity, only the i.i.d. case with the SP scenario is considered, as the results for other cases are similar. The observations are standardized with standard deviation of one. All the other settings are the same as those in Figure 1. For instance, the t distribution can be written as:

$$\begin{aligned}
&\boldsymbol{X}_j^{OC}|\boldsymbol{\mu},\boldsymbol{\theta} \sim t_3(\boldsymbol{I}_{m\times m}) + \boldsymbol{\mu}\\
&\mu_i|\theta_i \sim (1-\theta_i)\delta_0(\mu_i) + I(\theta_i=1)h_1(\mu_i) + I(\theta_i=-1)h_2(\mu_i)\\
&\theta_i \overset{i.i.d}{\sim} \text{Multinoulli}\,(p_0,p_1,p_{-1}),\ \sum_{k=0,1,-1} p_k = 1.
\end{aligned}$$

The simulation results of the four diagnostic procedures are summarized in Figure 7. From the plots, similar conclusions can be drawn as follows. The SOM procedure is always valid provided that n is not too small. The diagnostic performance of SDM is quite similar to that of SOM. MOW and MDW cannot control the MDR until n is large (e.g., $n \geq 15$). After the convergence, the EFP levels of MOW and MDW are significantly higher than those of SOM and SDM, respectively. Now, we can conclude that the proposed SOM and SDM procedures are still efficient without the assumption of normality.

5 | REAL-DATA ANALYSIS

In this section, we demonstrate the proposed data-driven procedure SDM by using a real-world dataset from the semiconductor manufacturing process (SMP). The dataset is available from the UCI Machine Learning Repository (http://archive.ics.uci.edu), and was automatically recorded by a computer system that manages the entire SMP from producer requests to laboratory analysis. It contains in total 1567 observation vectors, including 1463 conforming observations and 104 nonconforming observations. In each observation vector, the data dimension is 590, that is, there are 590 continuous measurements. For demonstration, we regard the 1463 conforming observations as the IC historical dataset and the 104 nonconforming observations as the collected OC data for fault classification.

First of all, exploratory data analysis is necessary. We begin by handling data streams that remain unchanged over time and those with missing data. After removing the constant or extremely discrete data streams, $m = 453$ data streams

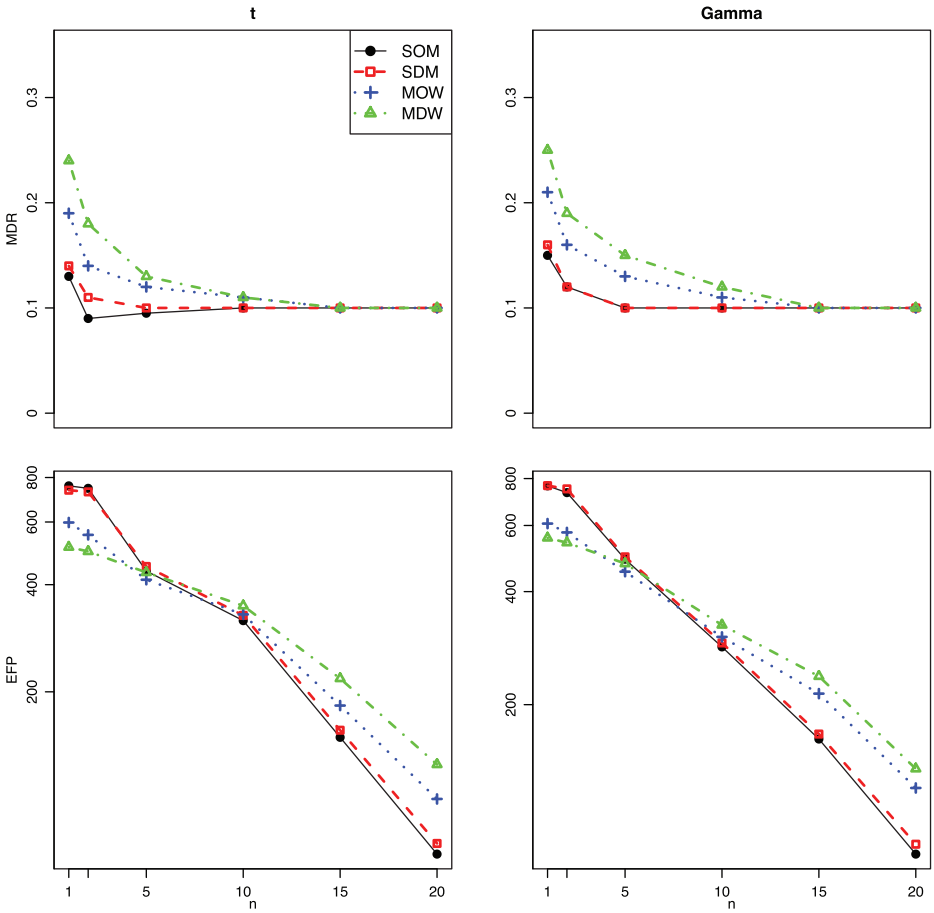

FIGURE 7 Diagnostic results of the t (the first column) and gamma (the second column) distributions for SOM (●), SDM (□), MOW (+), and MDW (▲). The MDR and EFP levels are respectively shown in the top and bottom rows

remain for further data analysis. Besides, we have found that the proportion of missing data is insignificant. Thus, we use the simple yet efficient mean imputation method to handle missing values. Then the Shapiro–Wilks GOF test is conducted, and the result implies that the distributions of many data streams are not normal. An inverse transformation, $\Phi^{-1}(\widehat{F}_k(X_{kt})), k = 1, \ldots, m$ is then implemented to approximately ensure the validity of the model assumption, where $\widehat{F}_k$ is the empirical c.d.f. of the kth data stream obtained from the 1463 IC observations and X_{kt} is the observed value of the kth data stream at time point t. We also calculate the sample correlation matrix and find that among the 102 378 lower triangular elements, there are 1098 elements with absolute values larger than 0.3, which is similar to the case of $\mathbf{\Sigma} = (\sigma_{ij}) = 0.4^{|i-j|}$. Therefore, we may conclude that a certain weak correlation structure exists.

To implement SDM, we still need to estimate important parameters in model (1), including the null proportion p_0, the p.d.f. of the shift size $h(\mu)$ and the marginal p.d.f. $g(x)$. First, we estimate p_0. Since the theoretical null marginal distribution is standard normal, we directly apply the method of Jin and Cai (2007) to estimate p_0. The corresponding estimate is denoted by $\widehat{p}_0$, and the estimation results are summarized in Table 2, from which it can be observed that $\widehat{p}_0$ becomes stable at around 0.6 as n gets larger. With $\widehat{p}_0$, we

TABLE 2 Estimated null proportion with various choices of n

n	10	30	60	90	104
$\widehat{p}_0$	0.66	0.64	0.58	0.60	0.61

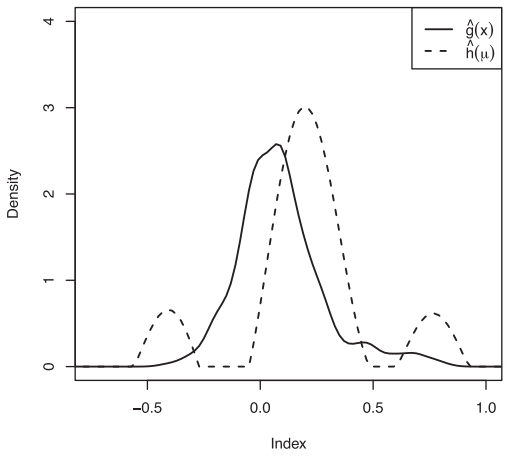

FIGURE 8 Illustration of $\widehat{g}(x)$ and $\widehat{h}(\mu)$

can estimate $h(\cdot)$ using the deconvoluting kernel estimator (7) discussed in Section 3.2. The resulting estimate is denoted by $\widehat{h}(\mu)$. We also estimate $g(x)$ by the classic kernel density estimate $\widehat{g}(x)$. For illustration, Figure 8 displays $\widehat{h}(\mu)$ and $\widehat{g}(x)$

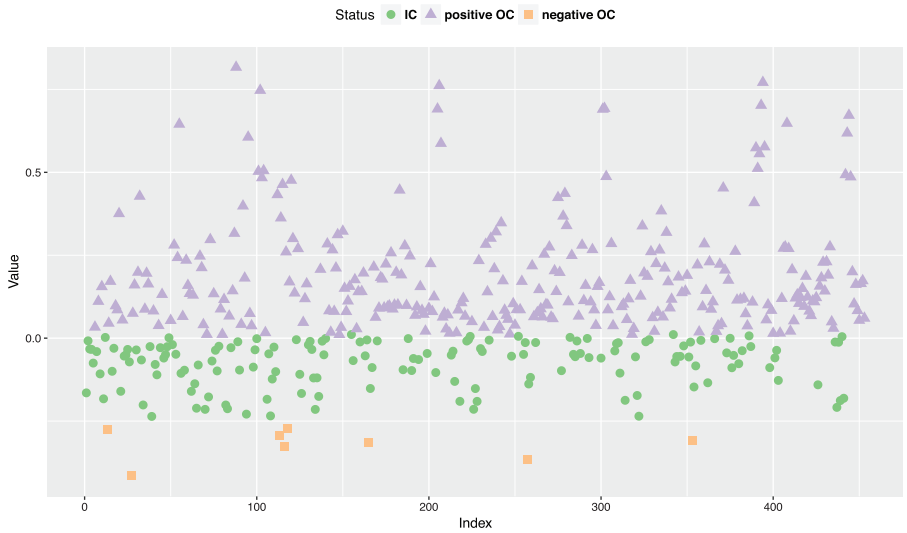

FIGURE 9 Diagnostic results of SDM with $\alpha = 0.1$ for the semiconductor manufacturing data

for cases when $n = 104$. It can be seen that both positive and negative shifts exist, and OC shifts focus on $\mu \in (0,0.5)$.

Now, we artificially use the nonconforming observations for fault classification, and apply the proposed SDM procedure to identify the OC data streams and determine the shift direction. Note that unlike simulation studies, it makes little sense to compare the diagnostic performance of SDM with its competitors in real-data analysis, since the actual MDR and EFP levels of the methods are unknown. Therefore, we only consider SDM here. With $\widehat{p}_0$, $\widehat{h}(\mu)$ and $\widehat{g}(x)$ discussed above, we can obtain $\widehat{H}_{max}(X_i)$, $\widehat{H}_0(X_i)$ and $\widehat{\Lambda}_i(X_i)$ and implement SDM immediately. The diagnostic result of SDM with $\alpha = 0.1$ is displayed in Figure 9, in which 298 data streams are deemed as OC. It can be expected that these streams includes 90% of the true OC streams with the right directions and few IC streams.

Before concluding, Figure 9 reveals something intriguing. It is well known in the field of multiple testing that applying the Lfdr-based procedure may cause observations located farther from the null distribution to be less significant than observations located closer to the null distribution when the non-null distribution is asymmetric about the null (Sun & Cai, 2007). Such a phenomenon can also be seen in Figure 9, in which more extreme observations may be accepted while less extreme observations may be rejected, and the upper rejection boundary is much closer to zero than the lower rejection boundary, leading to asymmetric rejection boundaries. The main reason is that $\widehat{h}(\mu)$ is large when $\mu \in (0,0.5)$, and thus SDM gives a higher priority to the observations in this interval.

6 | CONCLUDING REMARKS

In this article, we pioneer a novel formulation of the fault classification problem of HDDS on the basis of a three-classification multiple-testing framework. Based on a general parametric mixture model, an oracle diagnostic procedure for MDR-control and its data-driven version are proposed for finding the subset with the smallest EFP while controlling the MDR at given level α. The proposed data-driven procedure, SDM, is theoretically optimal, and computationally and statistically efficient. It enables more precise fault classification, and provides insights guiding further decisions, making it appealing in practice.

Many important issues remain that need to be solved. First, it is assumed in this paper that the change-point is detected correctly based on certain SPC methods (Pignatiello & Samuel, 2001; Zamba & Hawkins, 2006; Zou et al., 2007). However, in reality, the detected change-point may be inaccurate, or change-point detection may be coupled with fault classification. Greater research effort should be made on the whole post-signal process (including change-point detection and fault classification). Second, the proposed diagnostic procedure is designed only for OC mean shifts. In future research, a diagnostic procedure for OC variance–covariance shifts can be designed accordingly. Third, the proposed diagnostic procedure for MDR-control is designed from the perspective of mathematical expectation. However, the missed discovery percentage for a single execution may not be well controlled. It would be useful to develop a diagnostic procedure to ensure a high probability that the missed discovery percentage for each single execution is at least equal to a certain acceptable value. We leave this to future research. Fourth, the performance of the marginal procedure SDM may be affected seriously when the correlation is extremely strong or long-range. Therefore, diagnostic procedures that can effectively and efficiently utilize the correlation information must be developed in such cases. Finally, in many real world applications, the loss of missing a fault is likely different from the loss of missing its direction. It is important for us to

systematically investigate how to deal with both types of loss in the future.

ACKNOWLEDGMENTS

The authors sincerely acknowledge the efforts of the Editor, the Associate Editor, and two anonymous referees that have resulted in significant improvements of this paper. This paper is partly supported by National Natural Science Foundation of China (12071144, 71931004, 11871324, 11771145, 71931006, 11801210), RGC GRF ([16201718, 16216119), 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.

DATA AVAILABILITY STATEMENT

The dataset is available from the UCI Machine Learning Repository (http://archive.ics.uci.edu).

ORCID

Dongdong Xiang https://orcid.org/0000-0002-4609-7392
Fugee Tsung https://orcid.org/0000-0002-0575-8254
Yicheng Kang https://orcid.org/0000-0002-8512-2722

REFERENCES

Cai, T. T., & Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, *104*(488), 1467–1481.

Cai, T. T., Sun, W., & Wang, W. (2019). Cars: Covariate assisted ranking and screening for large-scale two-sample inference (with discussion). *Journal of the Royal Statistical Society, Series B*, *81*, 187–234.

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

Delaigle, A., & Gijbels, I. (2004). Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics and Data Analysis*, *45*(2), 249–267.

Delaigle, A., & Hall, P. (2006). On optimal kernel choice for deconvolution. *Statistics & Probability Letters*, *76*(15), 1594–1602.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, *99*(465), 96–104.

Holte, S. E., Lee, E. K., & Mei, Y. (2016). Symmetric directional false discovery rate control. *Statistical Methodology*, *33*, 71–82.

Jin, J., & Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, *102*(478), 495–506.

Li, J., Jin, J., & Shi, J. (2008). Causation-based t^2 decomposition for multivariate process monitoring and diagnosis. *Journal of Quality Technology*, *40*(1), 46–58.

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

Li, W., Zhang, C., Tsung, F., & Mei, Y. (2020). Nonparametric monitoring of multivariate data via knn learning. *International Journal of Production Research*, 1–16.

Lin, X., & Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*, *95*, 520–534.

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

Mason, R. L., Tracy, N. D., & Young, J. C. (1997). A practical approach for interpreting multivariate t^2 control chart signals. *Journal of Quality Technology*, *29*(4), 396–406.

May, G., & Spanos, C. (2006). Fundamentals of semiconductor manufacturing and process control. John Wiley & Sons.

Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika*, *97*(2), 419–433.

Pignatiello, J. J. J., & Samuel, T. R. (2001). Estimation of the change point of a normal process mean in spc applications. *Journal of Quality Technology*, *33*(1), 82–95.

Qiu, P., Zou, C., & Wang, Z. (2010). Nonparametric profile monitoring by mixed effects modeling. *Technometrics*, *52*(3), 265–277.

Silverman, B. W. (1986). Density estimation for statistics and data analysis (Vol. *26*). CRC Press.

Sullivan, J. H., Stoumbos, Z. G., Mason, R. L., & 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., & Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, *102*(479), 901–912.

Sun, W., & McLain, A. C. (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association*, *107*(498), 673–687.

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

Xian, X., Wang, A., & Liu, K. (2018). A nonparametric adaptive sampling strategy for online monitoring of big data streams. *Technometrics*, *60*(1), 14–25.

Xiang, D., Zhao, S., & Cai, T. T. (2019). Signal classification for the integrative analysis of multiple sequences of large-scale multiple tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*, 707–734.

Xie, J., Cai, T. T., Maris, J., & Li, H. (2011). Optimal false discovery rate control for dependent data. *Statistics and Its Interface*, *4*(4), 417–430.

Yan, H., Paynabar, K., & 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., & Hawkins, D. M. (2006). A multivariate change-point for statistical process control. *Technometrics*, *48*(4), 539–549.

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

Zhu, Y., & Jiang, W. (2009). An adaptive t^2 chart for multivariate process monitoring and diagnosis. *IIE Transactions*, *41*(11), 1007–1018.

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

Zou, C., & Qiu, P. (2009). Multivariate statistical process control using lasso. *Journal of the American Statistical Association*, *104*(488), 1586–1596.

Zou, C., Tsung, F., & Wang, Z. (2007). Monitoring general linear profiles using multivariate ewma schemes. *Technometrics*, *49*(4), 395–408.

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

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Xiang, D., Li, W., Tsung, F., Pu, X., & Kang, Y. (2021). Fault classification for high-dimensional data streams: A directional diagnostic framework based on multiple hypothesis testing. *Naval Research Logistics (NRL)*, *68*(7), 973–987. https://doi.org/10.1002/nav.22008