Simultaneous Optimal Control of Directional Missed Discovery Rates in Data Stream Diagnosis

Yan He
School of Statistics, East China Normal University, Shanghai, China

Yicheng Kang
Department of Information Systems & Analytics,
Miami University, Ohio, USA

Dongdong Xiang *
School of Statistics, East China Normal University, Shanghai, China

Peihua Qiu
Department of Biostatistics, The University of Florida

March 10, 2024

Abstract

High-dimensional data streams are ubiquitous in modern manufacturing because of their ability to provide valuable information about the industrial system's performance on a real-time basis. If a shift occurs in a production process, fault diagnosis based on the data streams is of critical importance for identifying the root cause. Existing methods have largely focused on controlling the total missed discovery rate without distinguishing missed signals for positive versus negative components of the shift vector. In practice, however, losses incurred from the two directional shifts can differ substantially, so it is desirable to constrain the proportions of missed signals for positive and negative components at two distinctive levels. In this article, we propose a fault classification procedure that controls the two proportions separately. By formulating the problem as Lagrangian multiplier optimization, we show that the proposed procedure is optimal in the sense that it minimizes the expected number of false discoveries. We also suggest an iterative adjustment algorithm that converges to the optimal Lagrangian parameters. The asymptotic optimality for the data-driven version of our procedure is established as well. Theoretical justification and numerical comparison with state-of-the-art methods show that the proposed procedure works well in applications.

Keywords: fault classification, high-dimensional data, large-scale testing, Markov models, post-signal diagnostics, quality control,

*__Corresponding author__: Dongdong Xiang, Email: terryxdd@163.com

1 Introduction

Modern manufacturing systems are often installed with large numbers of sensors that monitor a variety of process variables such as temperature, humidity, opacity, pressure, vibration and conductivity. These sensors continuously generate data in high volume at high velocity. Such data are often referred to as high-dimensional data streams (HDDS) in the literature. Because of their ability to offer real-time information about the industrial system's performance, HDDS are deemed to possess great potential for helping improve the practice of industrial quality control. In particular, HDDS can be used to achieve two objectives: quick delivery of out-of-control (OC) signals after the production process has had a shift (i.e., process monitoring) and accurate identification of the root cause after an OC signal has been given (i.e., fault diagnosis). Past research on HDDS analysis has mostly focused on process monitoring. See, for instance, Wang and Jiang (2009), Zou and Qiu (2009), Mei (2010), Capizzi and Masarotto (2011), Liu et al. (2015), Zou et al. (2015), Qiu (2018), Xian et al. (2018), Yan et al. (2018), Zhang et al. (2020), Li et al. (2021) and Kang (2022). Qiu (2020) and Woodall and Montgomery (2014) provided comprehensive overviews on this topic. More recently, there have been growing interests in post-signal diagnostic problems involving HDDS, in which the primary purpose is to identify the data streams affected by the OC event. The task of efficient fault diagnosis is particularly important in situations involving HDDS, as it would be practically impossible for quality engineers to examine thousands of data streams one by one. A number of diagnostic approachs have been proposed in the literature. Some have made use of variable selection techniques for multiple linear regression by setting up the fault diagnosis problem such that the selected variables are regarded as OC streams (e.g., Zou et al. 2011; Li et al. 2017; Ebrahimi et al. 2021). Although this approach has been shown to outperform many traditional methods in terms of isolating the OC streams, it is unable to determine the shift direction (positive or nega-

tive) in each component of the shift vector. In many applications, however, it is beneficial to have directional information, because it allows the quality engineer to shorten the diagnosis process and customize repairs. Others have analyzed the HDDS problem under the framework of large-scale multiple testing (e.g., Li et al. 2020; Xiang et al. 2021a,b). This approach enables inference about the shift directions and controls the proportion of OC streams missed by diagnosis (i.e., missed discovery rate, which is formally defined in Section 2). A major limitation of these methods is that they do not distinguish the missed OC streams (i.e. components) with positive shifts versus the missed OC streams with negative shifts. In some applications, however, potential risks associated with two directional shifts can differ substantially. For instance, the concentration of food preservatives, a carefully monitored process variable in food production, can reduce the food product's shelf life if the level of concentration is too low, resulting in possible food waste. On the other hand, too much preservatives have harmful side effects in form of headaches, allergies and cancer, jeopardizing public health (Sharma 2015). In other words, the implication of having too much preservatives is far more severe than that of having too little. This asymmetric decision-making situation makes it desirable to control the proportions of missed signals of two directional shifts separately at two distinctive levels.

In this article, we propose a data stream fault diagnosis procedure that controls proportions of missed OC components of two directional shifts at separate levels. Theoretically, by formulating the problem as a Lagrange multiplier optimization, we derive that our procedure minimizes the expected number of false discoveries. This property is particularly pertinent to manufacturing applications, as false discoveries would result in mistakenly discarding good-quality products. Therefore, our procedure not only identifies nearly all the OC streams but also excludes the irrelevant in-control (IC) streams. Numerically, it is challenging to find the optimal Lagrangian multipliers that correspond to the two pre-specified levels of missed discovery rates, because it involves solving two equations and the simple

thresholding method used in the single-constraint case (e.g., He et al. 2023) does not work in our multi-constraint scenario. To address this issue, we find that the missed discovery rates are monotonic in one Lagrange multiplier while fixing the other. It reflects the trade-off between the two constraints - aggressively controlling one directional missed discovery rate will inevitably loosen the other. We make use of this monotonic property and suggest an iterative adjustment algorithm that converges to the optimal Lagrange multipliers. The asymptotic optimality of the data-driven version of the proposed procedure is established as well. Simulation studies show that the data-driven version is almost as powerful as the oracle version, with superior performance over existing methods in various settings.

The remainder of this article is organized as follows. In Section 2, the proposed procedure and its theoretical properties are described in detail. In Section 3, the performance of our procedure is compared with state-of-the-art methods using simulations. Section 4 demonstrates the proposed procedure with a real manufacturing dataset. Section 5 summarizes the work and suggests future research directions. Computer code and technical proofs are included in the supplementary file.

2 Proposed Methodology

We organize the discussion of our methodology into three subsections. In Subsection 2.1, our fault classification problem is formulated as a three-way comparison and the optimal procedure is defined to be the solution to a related constrained optimization problem. In Subsection 2.2, the optimal procedure is derived under the assumption that all the model parameters are known. In Subsection 2.3, we propose methods for estimating the parameters and thus obtain the data-driven version of the optimal procedure.

2.1 Problem Formulation

Let $\boldsymbol{X}_t = (X_{1t}, X_{2t}, \ldots, X_{mt})^T$ denote the m process variables that are being monitored at time t. If the manufacturing process works stably, $\boldsymbol{X}_t$ follows an IC distribution, which can be estimated with an IC sample collected beforehand. The problem of IC distribution estimation is often considered in phase I statistical process control (SPC). We refer readers to Qiu (2014) and Montgomery (2020) for commonly used phase I SPC methods. We assume that the IC distribution is known throughout our discussion. After an OC event occurs in the production process, a control chart would signal the shift and a small number of OC observations, denoted by $\{\boldsymbol{X}_j^{\texttt{OC}}, j = 1, 2, \ldots, n\}$, are available for post-signal fault diagnosis. It is worth noting that not all the component of $\boldsymbol{X}_j^{\texttt{OC}}$ have shifted necessarily even though the process vector has been flagged OC. Knowing which components have shifted and their shift directions is critical to root cause identification. To this end, let θ_i be a hidden status variable associated with the i-th component (data stream) of $\boldsymbol{X}_j^{\texttt{OC}}$. The component remains IC if $\theta_i = 0$ and has shifted positively or negatively if $\theta_i = 1$ or -1. θ_i's are unobservable and we would like to infer their values based on $\{\boldsymbol{X}_j^{\texttt{OC}}\}$.

Specifically, we assume that $\{\boldsymbol{X}_j^{\texttt{OC}}\}$ are generated from the following mixture model:

$$
\begin{aligned}
&\boldsymbol{X}_j^{\texttt{OC}} \big| \, \boldsymbol{\mu} \sim F(\boldsymbol{x}_j \mid \boldsymbol{\mu}), \\
&\mu_i | \theta_i \sim (1 - |\theta_i|)\delta_0(\mu_i) + I(\theta_i = 1)h_1(\mu_i) + I(\theta_1 = -1)h_2(\mu_i), \\
&\theta_i \sim \pi_0\delta_0(\theta_i) + \pi_1\delta_1(\theta_i) + \pi_{-1}\delta_{-1}(\theta_i), \; i = 1, 2, \ldots, m, \; j = 1, 2, \ldots, n,
\end{aligned}
\tag{1}
$$

where $F(\cdot \mid \boldsymbol{\mu})$ denotes the joint conditional distribution of $\boldsymbol{X}_j^{\texttt{OC}}$ given $\boldsymbol{\mu}$, $\delta_x(\cdot)$ denotes the Dirac function with the unit mass concentrated at x, $I(\cdot)$ is an indicator function, $h_1(\cdot)$ and $h_2(\cdot)$ are probability density functions with support in $(0, +\infty)$ and $(-\infty, 0)$ respectively, and $\{\pi_k \geq 0, k = 0, \pm 1\}$ satisfy that $\sum_{k=-1}^{1} \pi_k = 1$. $\boldsymbol{\mu}$ is the OC mean with some nonzero components because of the process shift. θ_i represents the status of data stream i with the probability of being IC, having shifted in the positive direction

and having shifted in the negative direction equal to π_0, π_1 and π_{-1}, respectively. The hierarchical structure of model (1) has been widely used in the literature for approximating high-dimensional distributions (e.g., Efron 2004; Sun and Cai 2007; Cai and Sun 2009; Sun and Cai 2009). We further assume that $\{\boldsymbol{X}_j^{\texttt{OC}}\}$ are independent over time. In cases when the observations are temporally correlated, we can remove the correlation in advance by some decorrelation techniques (e.g., Apley and Tsung 2002; Qiu et al. 2020). Our oracle procedure in Subsection 2.2 does not require $F(\cdot \mid \boldsymbol{\mu})$ to be of any parametric form. The data-driven procedure in Subsection 2.3 further assumes that $F(\cdot \mid \boldsymbol{\mu})$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix Σ. This normality assumption is to ensure consistent estimation of the model parameters. In Subsection 3.2 and 3.3, we will consider cases when $\boldsymbol{\Sigma}$ is non-diagonal or the distribution of $\boldsymbol{X}_j^{\texttt{OC}}|\boldsymbol{\mu}$ is non-normal.

In many HDDS applications, OC streams tend to occur in clusters due to the fact that the physical locations of the corresponding sensors are installed in close vicinity. Moreover, the state of one variable would influence the state of the next variable. For instance, in automotive fault analysis, overheating detected by a thermal sensor can cause abnormally high vibration, which would be recorded by a vibration sensor (e.g., Bonnett and Soukup 1992; Randall 2021). Such correlations are informative for fault diagnosis. Notably, some data streams may remain completely IC during fault analysis after an OC signal is given. This is the case when some modules of an industrial system perform stably while some other modules have malfunctions. Hence it is possible for the state variable to transition from an abnormal state to the normal state as we examine the data streams. Based on the above motivations, we suggest the following three-state hidden Markov model (HMM). Specifically, assume that $\{\theta_i, i = 1, 2, \ldots, m\}$ form a stationary, irreducible and aperiodic Markov chain. Denote the initial distribution of the Markov chain by $\boldsymbol{\pi}^0 = (\pi_0^0, \pi_1^0, \pi_{-1}^0)^T$, i.e., $\theta_1 = \pi_k$ with probability π_k^0 for $k = 0, \pm 1$. Let the stationary distribution be $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_{-1})^T$, i.e., $\lim_{i\to\infty} P(\theta_i = k) = \pi_k$ for $k = 0, \pm 1$. Let $\mathcal{T} = \{a_{kl}, k, l = 0, \pm 1\}$ be the

transition matrix. That is,

$$a_{kl} = P\left(\theta_i = l | \theta_{i-1} = k\right), \; k, l = 0, \pm 1.$$

The transition probabilities do not depend on i and satisfy the standard constraints: $0 \leq a_{kl} \leq 1$, $k = 0, \pm 1$, and $\sum_{l=-1}^{1} a_{kl} = 1$ for $k \in \{0, \pm 1\}$. The stationary distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_{-1})^T$ can be determined by the transition matrix using the relationship $\boldsymbol{\pi}^T \mathcal{T} = \boldsymbol{\pi}^T$. In the special case of $\boldsymbol{\pi}^0 = \boldsymbol{\pi}$, we have $P(\theta_i = k) = \pi_k$. In the literature, HMMs have been used for change point detection in industrial diagnosis problems. See, for instance, Alippi et al. (2012) and Fuh and Mei (2015). It also has been shown that HMM-based change detection methods enjoy nice theoretical properties (Fuh and Tartakovsky, 2018). It is worth noting that these methods use HMMs to capture the temporal dynamics, i.e., to estimate the time point when the process has shifted. In our HDDS diagnosis problem, however, there is no temporal component as we focus on IC/OC classification at the data-stream level.

Our objective of post-signal diagnosis is three-fold: identify nearly all the OC streams, correctly specify the shift directions, and minimize the number of false discoveries. To achieve this objective, consider the following three-class testing problem:

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

Based on the observed values of $(\boldsymbol{X}_1^{\texttt{OC}}, \boldsymbol{X}_2^{\texttt{OC}}, \ldots, \boldsymbol{X}_n^{\texttt{OC}})$, we are interested in inferring the value of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^T$. A solution to problem (2) can be represented by a decision rule $\boldsymbol{d} : \Omega \to \{0, \pm 1\}^m$, where Ω is the sample space. Let $\boldsymbol{d} = (d_1, d_2, \ldots, d_m)^T$ be the decision vector of length m with elements 0, 1 or -1. If $d_i = 0$, then θ_i is regarded to have taken the value of 0, i.e., the i-th data stream is classified as IC. If $d_i = 1$ or -1, then the i-th data stream is flagged as OC with a positive shift or negative shift, respectively. For a given decision rule $\boldsymbol{d}$, the outcomes of problem (2) can be categorized as done in Table 1. Based on this categorization, some authors (e.g., Xiang et al. 2021a; He et al. 2023) have

Table 1: Classification of tested hypotheses

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

defined the total missed discovery rate (tMDR) as follows.

$$\text{tMDR}(\boldsymbol{d}) = \frac{\text{E}\left(N_{10} + N_{12} + N_{20} + N_{21}\right)}{\text{E}\left(m_1 + m_2\right)}.$$

It can be seen that tMDR is relevant to our diagnosis problem as controlling tMDR at a low level is equivalent to identifying almost all the OC streams. However, tMDR does not distinguish missed signals of positive shifts versus missed signals of negative shifts. For instance, it is possible to miss 10% signals of positive shifts and still achieve a 5% tMDR. This can cause the actual risk of missing a signal much higher than expected in some asymmetric decision-making situations (e.g., the food production example in Section 1). To address this limitation, define the two marginal missed discovery rates as follows.

$$\text{mMDR}_1(\boldsymbol{d}) = \frac{\text{E}\left(N_{10} + N_{12}\right)}{\text{E}\left(m_1\right)}, \quad \text{mMDR}_{-1}(\boldsymbol{d}) = \frac{\text{E}\left(N_{20} + N_{21}\right)}{\text{E}\left(m_2\right)}.$$

Instead of solely controlling tMDR, controlling $\text{mMDR}_{\pm 1}$ at their desired (and possibly different) levels leads to a more precise risk management for manufacturers.

In addition to identifying nearly all the OC streams, a reasonable fault diagnosis procedure should keep the expected number of false discoveries (EFD) to a minimum as it is wasteful to dispose of good-quality products. Therefore, we consider the following constrained optimization problem:

$$\min_{\boldsymbol{d}:\Omega \to \{0,\pm 1\}^m} \text{EFD}\left(\boldsymbol{d}\right) \text{ subject to } \text{mMDR}_1(\boldsymbol{d}) \leq \alpha_1 \text{ and } \text{mMDR}_{-1}(\boldsymbol{d}) \leq \alpha_{-1}, \tag{3}$$

where $\mathrm{EFD}(\boldsymbol{d}) = \mathrm{E}(N_{01} + N_{02})$, and $\alpha_{\pm 1} \in (0,1)$ are specified beforehand and represent possibly different levels of risk associated with missing the two directional signals. In the next subsection, we derive the solution to problem (3).

2.2 The Oracle Procedure

2.2.1 The Optimal Control of mMDRs

Throughout this subsection, we assume that the parameters $\mathcal{T}$, $\boldsymbol{\pi}^0$ and density functions $h_1(\cdot)$, $h_2(\cdot)$ in model (1) are known. Since the sample mean $\boldsymbol{X} = \sum_{j=1}^{n} \boldsymbol{X}_j^{\mathtt{OC}}/n$ is a sufficient statistic for $\boldsymbol{\mu}$, we will use it to construct our solution to problem (3).

Define

$$H_i^k(\boldsymbol{X}) = P(\theta_i = k|\boldsymbol{X}), \quad k = 0, \pm 1,\ i = 1, 2, \ldots, m.$$

Recall that a false discovery occurs if $\theta_i = 0$ and $d_i = \pm 1$. With iterative expectations, we can write

$$\begin{aligned}\mathrm{EFD}(\boldsymbol{d}) &= \mathrm{E}\left[\sum_{i=1}^{m} |d_i|\,(1 - |\theta_i|)\right] = \mathrm{E}\left\{\mathrm{E}\left[\sum_{i=1}^{m} |d_i|\,(1 - |\theta_i|)\,\middle|\,\boldsymbol{X}\right]\right\} \\ &= \mathrm{E}\left[\sum_{i=1}^{m}\sum_{k=\pm 1} I(d_i = k)\,P(\theta_i = 0|\boldsymbol{X})\right] = \mathrm{E}\left[\sum_{i=1}^{m}\sum_{k=\pm 1} I(d_i = k)\,H_i^0(\boldsymbol{X})\right].\end{aligned}$$

Similarly, we have

$$\mathrm{E}(N_{10} + N_{12}) = \mathrm{E}\left\{\sum_{i=1}^{m} [1 - I(d_i = 1)]\,H_i^1(\boldsymbol{X})\right\}, \tag{4}$$

$$\mathrm{E}(N_{20} + N_{21}) = \mathrm{E}\left\{\sum_{i=1}^{m} [1 - I(d_i = -1)]\,H_i^{-1}(\boldsymbol{X})\right\}, \tag{5}$$

$$\mathrm{E}(m_1) = \mathrm{E}\left[\sum_{i=1}^{m} H_i^1(\boldsymbol{X})\right], \quad \mathrm{E}(m_2) = \mathrm{E}\left[\sum_{i=1}^{m} H_i^{-1}(\boldsymbol{X})\right]. \tag{6}$$

Hence, the Lagrangian function for problem (3) is

$$L(\boldsymbol{\lambda}, \boldsymbol{d}) = \sum_{i=1}^{m} L_i(\boldsymbol{\lambda}, d_i) \overset{\mathtt{def}}{=} \sum_{i=1}^{m}\sum_{k=\pm 1} \left\{I(d_i = k)H_i^0(X) + \lambda_k\,[1 - I(d_i = k) - \alpha_k]\,H_i^k(\boldsymbol{X})\right\},$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_{-1})^T$ denotes the Lagrange multipliers. Given $\boldsymbol{\lambda}$, let us first consider the decision rule $d_i^{\boldsymbol{\lambda}}$ that minimizes $L_i(\boldsymbol{\lambda}, d_i)$. We have

$$L_i(\boldsymbol{\lambda}, 1) \leq L_i(\boldsymbol{\lambda}, 0) \iff H_i^0(\boldsymbol{X}) \leq \lambda_1 H_i^1(\boldsymbol{X}),$$
$$L_i(\boldsymbol{\lambda}, 1) \leq L_i(\boldsymbol{\lambda}, -1) \iff \lambda_{-1} H_i^{-1}(\boldsymbol{X}) \leq \lambda_1 H_i^1(\boldsymbol{X}).$$

It follows that $d_i^{\boldsymbol{\lambda}} = 1$ if $\max\{H_i^0(\boldsymbol{X}), \lambda_{-1} H_i^{-1}(\boldsymbol{X})\} \leq \lambda_1 H_i^1(\boldsymbol{X})$. We can similarly derive the conditions under which $d_i^{\boldsymbol{\lambda}} = -1$ and $d_i^{\boldsymbol{\lambda}} = 0$, respectively. Taken together, $L(\boldsymbol{\lambda}, \boldsymbol{d})$ is minimized by $\boldsymbol{d}^{\boldsymbol{\lambda}} = (d_1^{\boldsymbol{\lambda}}, d_2^{\boldsymbol{\lambda}}, \ldots, d_m^{\boldsymbol{\lambda}})^T$, defined as

$$d_i^{\boldsymbol{\lambda}} = \begin{cases} k, & \text{if } R_{k,i} \leq 0 \text{ and } R_{k,i} = \min\{R_{1,i},\ R_{-1,i}\} \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $R_{k,i} = H_i^0(\boldsymbol{X}) - \lambda_k H_i^k(\boldsymbol{X})$ and $k = \pm 1$. This is intuitively expected as $\boldsymbol{d}^{\boldsymbol{\lambda}}$ is based upon the comparison of the conditional probabilities $\{H_i^k(\boldsymbol{X}), k = 0, \pm 1\}$.

Next, we determine $\boldsymbol{\lambda}$ by solving

$$\text{mMDR}_k\left(\boldsymbol{d}^{\boldsymbol{\lambda}}\right) = \alpha_k, \quad k = \pm 1. \tag{8}$$

By (4) – (6), equation (8) is equivalent to

$$\text{E}\left\{\sum_{i=1}^{m}\left[1 - I(d_i^{\boldsymbol{\lambda}} = k) - \alpha_k\right] H_i^k(\boldsymbol{X})\right\} = 0, \quad k = \pm 1.$$

Let $N_k(\lambda_1, \lambda_{-1}) = \text{E}\left\{\left[1 - I(d_i^{\boldsymbol{\lambda}} = k) - \alpha_k\right] H_i^k(\boldsymbol{X})\right\}$, $k = \pm 1$. The solution to (8) can be obtained if we can find $\boldsymbol{\lambda}$ such that

$$N_k(\lambda_1, \lambda_{-1}) = 0, \quad , k = \pm 1. \tag{9}$$

It can be shown that $N_k(\lambda_1, \lambda_{-1})$ is decreasing in λ_k and increasing in $\lambda_{k'}$, where $k' \neq k$ (the proof is given in the supplementary file). Making use of this monotonic property, we can solve (9) by iteratively adjusting λ_1 and λ_{-1}. The next proposition formalizes this argument.

Proposition 1. *Let* $\check{\lambda}_{1,0}$ *and* $\check{\lambda}_{-1,0}$ *be initial values for* λ_1 *and* λ_{-1} *respectively. Given* $\check{\lambda}_{1,t-1}$ *and* $\check{\lambda}_{-1,t-1}$*, for* $t \geq 1$*, obtain* $\check{\lambda}_{1,t}$ *and* $\check{\lambda}_{-1,t}$ *as follows.*

$$\check{\lambda}_{1,t} = \sup\left\{\lambda \geq \check{\lambda}_{1,t-1} : N_1(\lambda, \check{\lambda}_{-1,t-1}) \geq 0\right\}, \tag{10}$$

$$\check{\lambda}_{-1,t} = \sup\left\{\lambda \geq \check{\lambda}_{-1,t-1} : N_{-1}(\check{\lambda}_{1,t-1}, \lambda) \geq 0\right\}. \tag{11}$$

Assume that there exists sufficiently large $A_{\pm 1} > 0$ *such that* $N_{\pm 1}(A_1, A_{-1}) < 0$*. If* $\alpha_1 + \alpha_{-1} \leq 1$ *and* $\check{\lambda}_{\pm 1,0} = 0$*, then sequences* $\{\check{\lambda}_{1,t}, t \geq 0\}$ *and* $\{\check{\lambda}_{-1,t}, t \geq 0\}$ *both converge. Furthermore,* $N_k(\lambda_1^*, \lambda_{-1}^*) = 0$ *for* $k = \pm 1$*, where* $\lambda_{\pm 1}^* = \lim_{t\to\infty} \lambda_{\pm 1,t}$*.*

Proposition 1 shows that the updating algorithm (10) – (11) converges to the solution to (8). The condition $N_{\pm 1}(A_1, A_{-1}) < 0$ is not restrictive. Suppose that the penalties for missing a signal of either direction are extremely large. Then the decision rule would never claim $\theta_i = 0$. Instead, it will determine θ_i to be 1 or -1 based on the comparison of $H_i^1(\boldsymbol{X})$ and $H_i^{-1}(\boldsymbol{X})$. The missed discovery rates should be very low in such an extreme case, thus complying with the constraints $\text{mMDR}_{\pm 1} < \alpha_{\pm 1}$. In other words, the pre-specified values of $\alpha_{\pm 1}$ can not be too small in order for the constrained optimization problem (3) to have a solution. The condition $\alpha_1 + \alpha_{-1} \leq 1$ is also mild. Commonly used values in practice (e.g., $\alpha_{\pm 1} = 0.1$) satisfy this condition. With $\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_{-1}^*)^T$ given in Proposition 1, the next theorem establishes the optimality of the decision rule $\boldsymbol{d}^{\boldsymbol{\lambda}^*}$. It shows that the optimality is achieved by spending all the available MDRs. Similar trade-offs exist in the multiple testing literature (e.g., Cai and Sun 2009).

Theorem 1. *If the conditions in Proposition 1 hold, then we have*

1. $\text{mMDR}_k\left(\boldsymbol{d}^{\boldsymbol{\lambda}^*}\right) = \alpha_k$*,* $k = \pm 1$*.*

2. *for any decision rule* $\boldsymbol{d}$ *satisfying* $\text{mMDR}_k \leq \alpha_k$ *for* $k = \pm 1$*,* $\text{EFD}\left(\boldsymbol{d}^{\boldsymbol{\lambda}^*}\right) \leq \text{EFD}\left(\boldsymbol{d}\right)$*.*

The updating algorithm (10) – (11) involves a series of expectations and suprema, whose calculations can be nontrivial. We approximate the expectation involved in $N_k(\lambda_1, \lambda_{-1})$ by

its moment estimator

$$N_k^*(\lambda_1, \lambda_{-1}) = \frac{1}{m}\sum_{i=1}^{m}\left[1 - I(d_i^{\boldsymbol{\lambda}}) - \alpha_k\right] H_i^k(\boldsymbol{X}).$$

Next, we find the updates based on N_k^*. Given $\lambda_1^{\texttt{old}}$ and $\lambda_{-1}^{\texttt{old}}$, let us first consider updating λ_1. Let $\eta_i^{\lambda_{-1}^{\texttt{old}}}(\boldsymbol{X}) = \max\{\lambda_{-1}^{\texttt{old}} H_i^{-1}(\boldsymbol{X}), H_i^0(\boldsymbol{X})\}/H_i^1(\boldsymbol{X})$. Write

$$\begin{aligned} N_1^*(\lambda_1, \lambda_{-1}^{\texttt{old}}) \geq 0 &\iff \sum_{i=1}^{m} I\left(\lambda_1 \geq \eta_i^{\lambda_{-1}^{\texttt{old}}}(\boldsymbol{X})\right) H_i^1(\boldsymbol{X}) \leq (1-\alpha_1)\sum_{i=1}^{m} H_i^1(\boldsymbol{X}) \\ &\iff \frac{\sum_{i=1}^{m} I\left(\lambda_1 \geq \eta_i^{\lambda_{-1}^{\texttt{old}}}(\boldsymbol{X})\right) H_i^1(\boldsymbol{X})}{\sum_{i=1}^{m} H_i^1(\boldsymbol{X})} \leq 1-\alpha_1. \end{aligned}$$

Therefore, we update λ_1 by

$$\lambda_1^{\texttt{new}} = \max\left\{\eta_l^{\lambda_{-1}^{\texttt{old}}}(\boldsymbol{X}) \geq \lambda_1^{\texttt{old}} : \frac{\sum_{i=1}^{m} I\left(\eta_l^{\lambda_{-1}^{\texttt{old}}}(\boldsymbol{X}) \geq \eta_i^{\lambda_{-1}^{\texttt{old}}}(\boldsymbol{X})\right) H_i^1(\boldsymbol{X})}{\sum_{i=1}^{m} H_i^1(\boldsymbol{X})} \leq 1-\alpha_1,\ l = 1, \ldots, m\right\}.$$

Similarly, we update λ_{-1} by

$$\lambda_{-1}^{\texttt{new}} = \max\left\{\xi_l^{\lambda_1^{\texttt{old}}}(\boldsymbol{X}) \geq \lambda_{-1}^{\texttt{old}} : \frac{\sum_{i=1}^{m} I\left(\xi_l^{\lambda_1^{\texttt{old}}}(\boldsymbol{X}) \geq \xi_i^{\lambda_1^{\texttt{old}}}(\boldsymbol{X})\right) H_i^{-1}(\boldsymbol{X})}{\sum_{i=1}^{m} H_i^{-1}(\boldsymbol{X})} \leq 1-\alpha_{-1},\ l = 1, \ldots, m\right\},$$

where $\xi_i^{\lambda_1} = \max\{\lambda_1 H_i^1(\boldsymbol{X}), H_i^0(\boldsymbol{X})\}/H_i^{-1}(\boldsymbol{X})$.

2.2.2 Computational Considerations

The use of decision rule $\boldsymbol{d}^{\lambda^*}$ requires the calculation of $H_i^k(\boldsymbol{X})$, which can be greatly simplified by the forward-backward algorithm described below. Let $f(\cdot)$ and $f(\cdot|\cdot)$ denote the joint and conditional likelihood, respectively. Define $\alpha_{i+1}(k) = f(X_1, X_2, \ldots, X_{i+1}, \theta_{i+1} = k)$ and $\beta_i(k) = f(X_{i+1}, X_{i+2}, \ldots, X_m|\theta_i = k)$. By using the Markov property repeatedly, we

have the following recursive computation.

$$\begin{aligned}\alpha_{i+1}(k) &= f(X_{i+1}|X_1, X_2, \ldots, X_i, \theta_{i+1} = k) f(X_1, X_2, \ldots, X_i, \theta_{i+1} = k) \\ &= f(X_{i+1}|\theta_{i+1} = k) \sum_{l=-1}^{1} f(X_1, X_2, \ldots, X_i, \theta_i = l, \theta_{i+1} = k) \\ &= f(X_{i+1}|\theta_{i+1} = k) \sum_{l=-1}^{1} f(X_1, X_2, \ldots, X_i, \theta_i = l) f(\theta_{i+1} = k|X_1, X_2, \ldots, X_i, \theta_i = l) \\ &= f(X_{i+1}|\theta_{i+1} = k) \sum_{l=-1}^{1} \alpha_i(l) a_{lk}. \end{aligned} \tag{12}$$

Similarly,

$$\beta_i(k) = \sum_{l=-1}^{1} a_{kl} \beta_{i+1}(l) f(X_{i+1}|\theta_{i+1} = l). \tag{13}$$

For initialization, we set $\alpha_1(k) = \pi_k^0 f(X_1|\theta_1 = k)$ and $\beta_m(k) = 1$. Taken together, we have

$$\begin{aligned} f(\boldsymbol{X}, \theta_i = k) &= f(X_1, \ldots, X_i, \theta_i = k) f(X_{i+1}, \ldots, X_m|X_1, \ldots, X_i, \theta_i = k) = \alpha_i(k)\beta_i(k), \\ H_i^k(\boldsymbol{X}) &= \frac{f(\boldsymbol{X}, \theta_i = k)}{f(\boldsymbol{X})} = \frac{\alpha_i(k)\beta_i(k)}{\alpha_i(0)\beta_i(0) + \alpha_i(1)\beta_i(1) + \alpha_i(-1)\beta_i(-1)}, \end{aligned} \tag{14}$$

We now have all the ingredients for our oracle procedure, which is summarized below.

The Oracle Procedure

1. Compute $\{\alpha_i(k), \beta_i(k), i = 1, \ldots, m, k = 0, \pm 1\}$ by recursive formula (12) – (13).

2. Compute $\{H_i^{\pm 1}(\boldsymbol{X}), H_i^0(\boldsymbol{X})\}$ by (14).

3. Run the algorithm (10) – (11) and denote the limit of the sequence $\{\check{\boldsymbol{\lambda}}_t\}$ by $\boldsymbol{\lambda}^*$.

4. Obtain the decision rule $\boldsymbol{d}^{\boldsymbol{\lambda}^*}$ according to (7).

It can be checked that the computational complexity in steps 1, 2 and 4 is $O(m)$. As for step 3, one iteration in the updating algorithm (10) – (11) is of computational complexity of $O(m)$. Since the algorithm is guaranteed to converge in a given precision, it takes a finite number of iterations to converge. Therefore, step 3 is also of computational complexity of $O(m)$, and the amount of computation involved in the entire procedure is $O(m)$. This conclusion will be confirmed by the numerical analysis in Subsection 3.4.

2.3 The Data-Driven Procedure

The model parameters $\{\mathcal{T}, \boldsymbol{\pi}^0\}$ and density functions $\{h_1(\cdot), h_2(\cdot)\}$ are rarely known in practice. In this subsection, we suggest ways for estimating them and thus obtain a data-driven version of our diagnostic procedure. To ensure that our estimates are consistent, we further assume that the conditional distribution of $\boldsymbol{X}_j^{\mathsf{OC}}$ given $\boldsymbol{\mu}$ is a multivariate normal distribution, $N(\mu, \Sigma)$, and Σ is the $m \times m$ identity matrix.

We first consider the estimation of $\{h_1(\cdot), h_2(\cdot)\}$. By the data generating mechanism given in model (1), we can write

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

where $\mu_i \sim \pi_0\delta_0(\mu_i) + \pi_1 h_1(\mu_i) + \pi_{-1} h_2(\mu_i)$, $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma}/n)$, and $\boldsymbol{\mu}$ and $\boldsymbol{\varepsilon}$ are independent. Given the observed value of $\boldsymbol{X}$, we are interested in estimating the density of μ_i. This is a typical setup for density deconvolution problems (e.g., Diggle and Hall 1993; Hall and Qiu 2005; Yi et al. 2021). It has been shown in the literature that deconvoluting kernel estimators (e.g., Fan 1991; Meister 2009) enjoy nice theoretical properties. Its idea is briefly described next. Let Ψ, Ψ_μ and Ψ_ε denote the characteristic function of X_i, μ_i and ε_i, respectively. We have $\Psi = \Psi_\mu \Psi_\varepsilon$. The kernel estimator for $\Psi(t)$ is $\widehat{\Psi}_X(t)\Psi_K(\tau t)$, where $\widehat{\Psi}_X(t) = 1/m \sum_{i=1}^m e^{\sqrt{-1}tX_i}$ is the empirical characteristic function, $\Psi_K(\cdot)$ is the Fourier transform of the kernel function $K(\cdot)$, and τ is the bandwidth parameter. Then the density of μ_i can be estimated by the inverse Fourier transform of $\widehat{\Psi}_X(\cdot)\Psi_K(\tau\cdot)/\Psi_\varepsilon(\cdot)$. Specifically, define

$$h(\mu) = \frac{\pi_1}{1-\pi_0} h_1(\mu) + \frac{\pi_{-1}}{1-\pi_0} h_2(\mu).$$

A natural estimator for $h(\mu)$ is given by

$$\widehat{h}(\mu) = \frac{1}{2\pi(1-\widehat{\pi}_0)} \int_{-\infty}^{\infty} e^{-\sqrt{-1}t\mu} \left[\frac{\widehat{\Psi}(t)}{\Psi_\varepsilon(t)} - \widehat{\pi}_0 \right] \Psi_K(\tau t)\, dt, \tag{15}$$

where $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_0, \widehat{\pi}_1, \widehat{\pi}_{-1})^T$ denotes the estimated stationary probabilities. In the literature, the estimator for $\boldsymbol{\pi}$ proposed in Jin and Cai (2007) has been shown to be consistent and work well in large-scale hypothesis testing and thus we adopt their estimator here. Since h_1 and h_2 have support in $(0, \infty)$ and $(-\infty, 0)$ respectively, we can estimate them by $\widehat{h}_1(\mu) = I(\mu > 0)\widehat{h}(\mu)(1 - \widehat{\pi}_0)/\widehat{\pi}_1$ and $\widehat{h}_2(\mu) = I(\mu < 0)\widehat{h}(\mu)(1 - \widehat{\pi}_0)/\widehat{\pi}_{-1}$. The choice of the kernel function K and bandwidth τ is crucial for the success of deconvoluting kernel estimators. In our implementation in Section 3 and Section 4, we adopt the sinc kernel $K(x) = \sin(x)/(\pi x)$ suggested by Delaigle and Hall (2006). As for τ, we select its value using the approach proposed in Delaigle and Gijbels (2004). The idea of using density deconvolution techniques for estimating h was initially proposed in Sun and McLain (2012), where they considered the i.i.d. two-class case. Notably, the correlation among $\{\mu_i\}$ has no influence on the optimal bandwidth choice and optimal rate of mean squared error for the kernel estimator in cases where the distribution of ε_i is normal (i.e., supersmooth). See Kulik (2008) for a detailed discussion about deconvolution and dependence. Therefore, we can use the deconvoluting estimator as if $\{\mu_i\}$ were independent.

Next, we estimate $\{f(x_i|\theta_i = k), k = 0, \pm 1\}$. It is clear that $f(x_i|\theta_i = 0)$ is the density of $N(0, 1/n)$. As for $\{f(x_i|\theta_i = k), k = \pm 1\}$, write

$$f(x_i|\theta_i = 1) = \int_0^\infty f(x_i, \mu_i|\theta_i = 1)\, d\mu_i = \int_0^\infty f(x_i|\mu_i, \theta_i = 1)h_1(\mu_i)\, d\mu_i,$$

where $f(x_i|\mu_i, \theta_i = 1)$ is the density of $N(\mu_i, 1/n)$. So we estimate $f(x_i|\theta_i = 1)$ by

$$\widehat{f}(x_i|\theta_i = 1) = \int_0^\infty f(x_i|\mu_i, \theta_i = 1)\widehat{h}_1(\mu_i)\, d\mu_i.$$

Similarly,

$$\widehat{f}(x_i|\theta_i = -1) = \int_{-\infty}^0 f(x_i|\mu_i, \theta_i = -1)\widehat{h}_2(\mu_i)\, d\mu_i.$$

Next, we describe our procedure for estimating the transition matrix $\mathcal{T}$ and initial distribution $\boldsymbol{\pi}^0$. Since $\{\theta_i\}$ are unobservable, the maximum likelihood estimators for $\mathcal{T}$

and $\boldsymbol{\pi}^0$ are not readily available. The expectation-maximization (EM) algorithm is useful in such a situation. The key step is that, given the values of $\mathcal{T}^{(t-1)}$ and $\boldsymbol{\pi}^{0(t-1)}$ at the t-th iteration of the EM loop, we can update them by

$$\pi_k^{0(t)} = P\left(\theta_1 = k|\boldsymbol{X}, \boldsymbol{\pi}^{0(t-1)}, \mathcal{T}^{(t-1)}\right),$$
$$a_{kl}^{(t)} = \frac{\sum_{i=1}^{m-1} P(\theta_i = k, \theta_{i+1} = l|\boldsymbol{X}, \mathcal{T}^{(t-1)})}{\sum_{i=1}^{m-1} P(\theta_i = k|\boldsymbol{X}, \boldsymbol{\pi}^{0(t-1)}, \mathcal{T}^{(t-1)})}, \ k, l = 0, \pm 1.$$

The full description of the EM algorithm is given below.

The EM Algorithm

1. Let $\{\mathcal{T}^{(0)}, \boldsymbol{\pi}^{0(0)}\}$ be the randomly initialized values.

2. At the t-th iteration, $t = 1, 2, \ldots,$

   (a) (E-step) compute

   - $\{\alpha_i^{(t-1)}(k), \beta_i^{(t-1)}(k), i = 1, 2, \ldots, m, k = 0, \pm 1\}$ given $\widehat{f}(x_i|\theta_i)$, $\mathcal{T}^{(t-1)}$ and $\boldsymbol{\pi}^{0(t-1)}$ using the recursive formula (12) - (13);
   - $P(\theta_i = k|\boldsymbol{X}, \boldsymbol{\pi}^{0(t-1)}, \mathcal{T}^{(t-1)}) = \frac{\alpha_i^{(t-1)}(k)\beta_i^{(t-1)}(k)}{\sum_{\xi=-1}^{1} \alpha_i^{(t-1)}(\xi)\beta_i^{(t-1)}(\xi)}$;
   - $P(\theta_i = k, \theta_{i+1} = l|\boldsymbol{X}, \boldsymbol{\pi}^{0(t-1)}, \mathcal{T}^{(t-1)}) = \frac{\alpha_i^{(t-1)}(k)a_{kl}^{(t-1)}\widehat{f}(x_{i+1}|\theta_{i+1}=l)\beta_{i+1}^{(t-1)}(l)}{\sum_{\xi=-1}^{1} \alpha_i^{(t-1)}(\xi)\beta_i^{(t-1)}(\xi)}$.

   (b) (M-step) update the parameters by

   - $\pi_k^{0(t)} = P(\theta_1 = k|\boldsymbol{X}, \boldsymbol{\pi}^{0(t-1)}, \mathcal{T}^{(t-1)}) = \frac{\alpha_1^{(t-1)}(k)\beta_1^{(t-1)}(k)}{\sum_{\xi=-1}^{1} \alpha_1^{(t-1)}(\xi)\beta_1^{(t-1)}(\xi)}$.
   - $a_{kl}^{(t)} = \frac{\sum_{i=1}^{m-1} P(\theta_i=k,\theta_{i+1}=l|\boldsymbol{X},\boldsymbol{\pi}^{0(t-1)},\mathcal{T}^{(t-1)})}{\sum_{i=1}^{m-1} P(\theta_i=k|\boldsymbol{X},\boldsymbol{\pi}^{0(t-1)},\mathcal{T}^{(t-1)})}$.

Denote the converging limit of the EM algorithm by $\widehat{\mathcal{T}}$ and $\widehat{\boldsymbol{\pi}}^0$. Based on these estimated values, $\widehat{H}_i^k(\boldsymbol{X})$ can be obtained according to (14). We can also estimate $N_k(\lambda_1, \lambda_{-1})$ by

$$\widehat{N}_k(\lambda_1, \lambda_{-1}) = \frac{1}{m}\sum_{i=1}^{m}\left\{\left[1 - I\left(\widehat{d}_i^{\boldsymbol{\lambda}} = k\right) - \alpha_k\right]\widehat{H}_i^k(\boldsymbol{X})\right\},$$

where $\widehat{\boldsymbol{d}}^{\boldsymbol{\lambda}}$ is given by (7) with $H_i^k(\boldsymbol{X})$ replaced by $\widehat{H}_i^k(\boldsymbol{X})$. Finally, we estimate $\boldsymbol{\lambda}^*$ by the updating algorithm (10) – (11) based on $\widehat{N}_k(\lambda_1, \lambda_{-1})$. We refer to $\widehat{\boldsymbol{d}}^{\widehat{\boldsymbol{\lambda}}^*}$ as the data-driven procedure. The next theorem shows that $\widehat{\boldsymbol{d}}^{\widehat{\boldsymbol{\lambda}}^*}$ is asymptotically optimal.

Theorem 2. *Assume that the conditions in Theorem 1 and the regularity conditions in the supplementary file hold. We have*

1. $\mathrm{mMDR}_k\left(\widehat{\boldsymbol{d}}^{\widehat{\boldsymbol{\lambda}}^*}\right) = \alpha_k + o(1),\ k = \pm 1;$

2. $\mathrm{EFD}\left(\widehat{\boldsymbol{d}}^{\widehat{\boldsymbol{\lambda}}^*}\right)/\mathrm{EFD}\left(\boldsymbol{d}^{\boldsymbol{\lambda}^*}\right) = 1 + o(1).$

3 Numerical Studies

We assess the numerical performance of our procedure in this section. Denote the oracle and data-driven versions of the proposed procedure by Oracle and Data-driven, respectively. In the recent fault analysis literature, Li et al. (2020) have proposed an MDR-based fault classification method. It determines the shift directions using the signs of observations. Xiang et al. (2021a) have also developed an MDR-based fault classification method and the authors have shown that their method is optimal in controlling the tMDR provided that the data streams are independent. We use these two methods as benchmarks in our comparison. Only the oracle versions of the two methods are considered. Denote the method in Li et al. (2020) as LO and the method in Xiang et al. (2021a) as XO.

Throughout this section, m is set equal to 3000, $\boldsymbol{\pi}^0$ is chosen to be $(1, 0, 0)^T$, and the following repeated simulations are done for each procedure. After $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ are given, the actual MDR and EFD values of each procedure are calculated based on 100 replicated simulations of $\boldsymbol{X}$. This whole process is then repeated 100 times, rendering 100 sets of $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ values along with 100 pairs of MDR and EFD values. The averages of these 100 MDR and EFD values are reported as the final metrics. We consider three scenarios: (i) $\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\theta}$ is normal with diagonal $\boldsymbol{\Sigma}$, (ii) $\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\theta}$ is normal with non-diagonal $\boldsymbol{\Sigma}$, and (iii) $\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\theta}$ is non-normal. In all settings, we consider $\alpha_{\pm 1} = 0.1$ and $\alpha_1 = 0.1$, $\alpha_{-1} = 0.05$. Since neither LO nor XO controls the directional MDRs, we have their the nominal tMDR level, α, induced by $\alpha_{\pm 1}$ for fair comparison. It follows from the definition that $\alpha = \alpha_{\pm 1}$ if

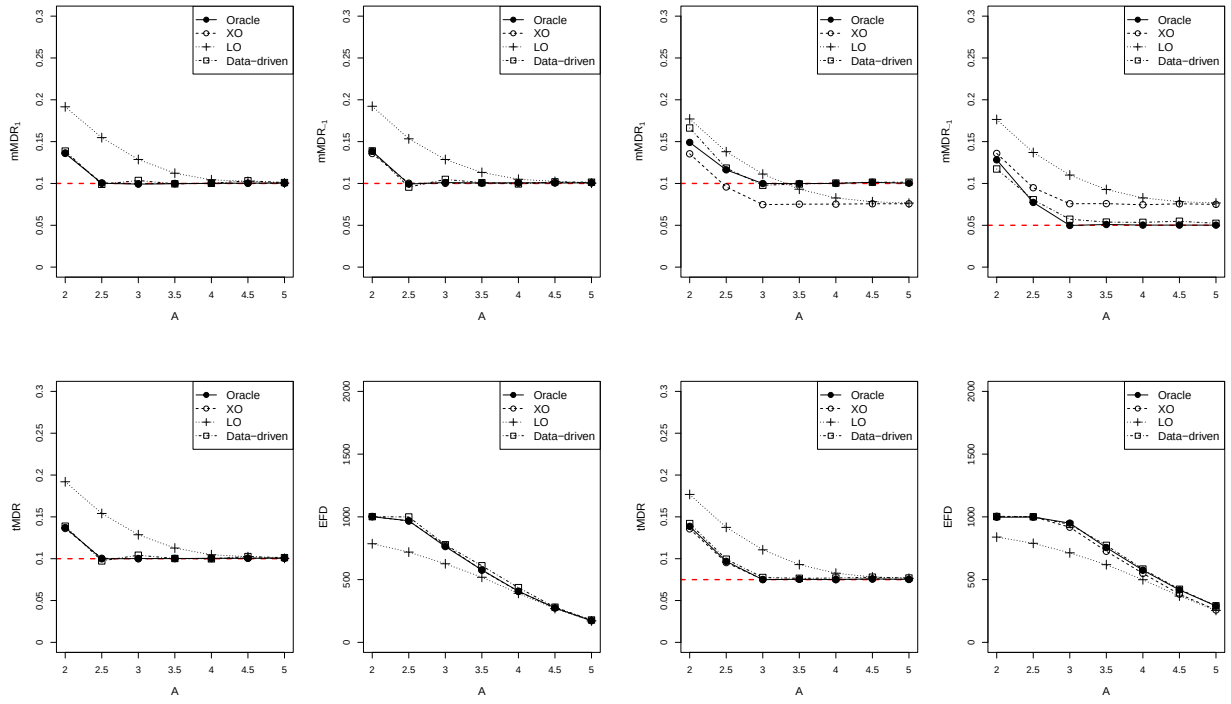
$\alpha_1 = \alpha_{-1}$. In cases where $\alpha_1 \neq \alpha_{-1}$, the induced value is $\alpha = \mathrm{E}(\alpha_1 m_1 + \alpha_{-1} m_2)/\mathrm{E}(m_1 + m_2)$ with m_1 and m_2 as defined in Table 1. Regarding the choice of $\{h_1(\cdot), h_2(\cdot)\}$, we consider $h_s(\cdot) = \mathrm{Gamma}(A_s, B_s, C_s)$ where $s = 1, 2$ and $\mathrm{Gamma}(a, b, c)$ denotes the density of the gamma distribution with shape parameter a, location parameter b and scale parameter c. The use of a location parameter here is to ensure that the magnitude of OC signals is bounded from below by a positive number.

3.1 Normal Cases with Diagonal Σ

We first evaluate the impact of shift size on the performance of our procedure. Let $h_1(\mu) = h_2(-\mu) = \mathrm{Gamma}(A, 0.05, 0.5)$ where $A \in (2, 5)$. Hence, the average shift size ranges from 1.05 to 2.55. Let $n = 2$ and

$$\mathcal{T} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}.$$

This transition matrix implies that $\{\theta_i, i = 1, \ldots, m\}$ are independent. The numerical results for this scenario are summarized in Figure 1. It can be seen from the figure that (i) all procedures perform better as the shift size increases, (ii) LO and XO could not control the $\mathrm{mMDR}_{\pm 1}$ at their nominal levels in the case of $\alpha_1 \neq \alpha_{-1}$, (iii) the proposed procedures (Oracle and Data-driven) are able to constrain the $\mathrm{mMDR}_{\pm 1}$ at the desired levels in both settings, and (iv) the proposed procedures are comparable with XO in the case of $\alpha_1 = \alpha_{-1}$. It is worth noting that the nominal tMDR is the same as the nominal $\mathrm{mMDR}_{\pm 1}$ if $\alpha_1 = \alpha_{-1}$ and XO is optimal in such a case. The simulation results show that, in addition to outperforming the other methods in the case of $\alpha_1 \neq \alpha_{-1}$, our procedures achieve the optimal performance if we are only concerned with controlling tMDR in the case of independent data streams.



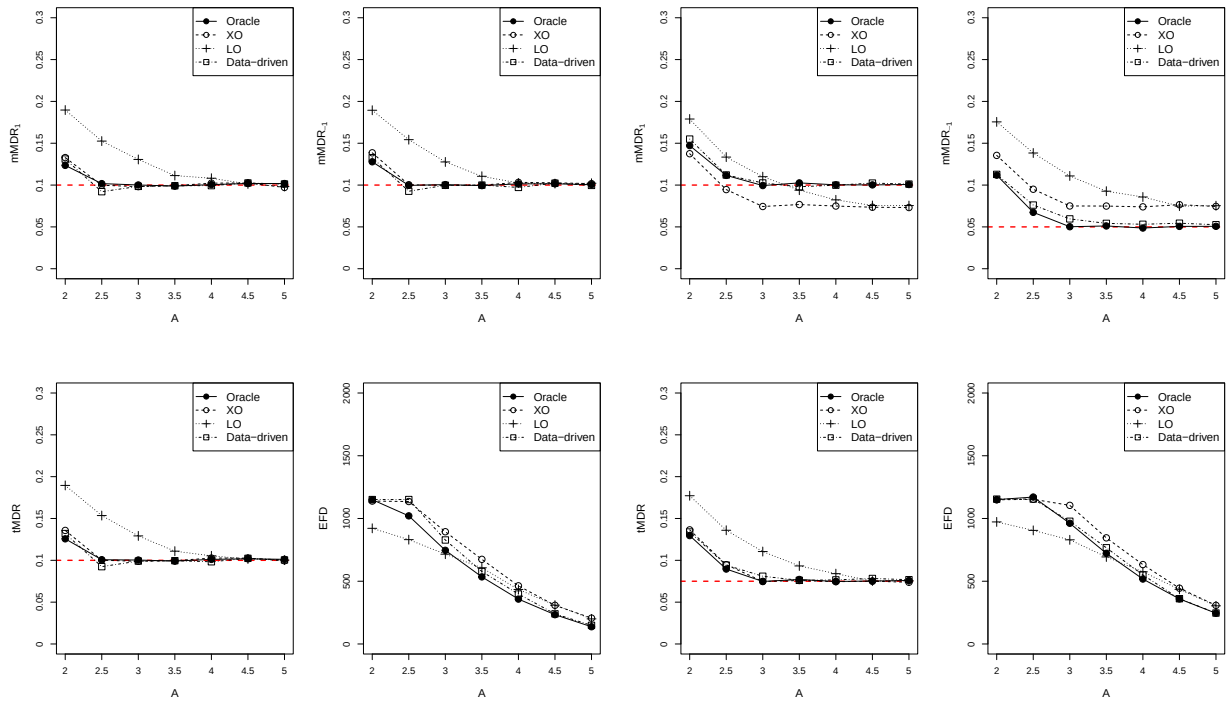
(a) $\alpha_1 = \alpha_{-1} = 0.1$ (b) $\alpha_1 = 0.1,\ \alpha_{-1} = 0.05$

Figure 1: Numerical comparison with LO and XO in the case where θ_i's are independent and the shift size varies from 1.05 to 2.55.

In the same setup as above, consider

$$\mathcal{T} = \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}.$$

This transition matrix introduces moderate correlation among θ_i's. The comparison results are shown in Figure 2. We see that Oracle and Data-driven are able to control $\text{mMDR}_{\pm 1}$ at the given levels in both settings and outperform XO in terms of EFD. The mMDR_{-1} values of LO and XO exceed the pre-specified level in the case of $\alpha_1 = 0.1$ and $\alpha_{-1} = 0.05$.



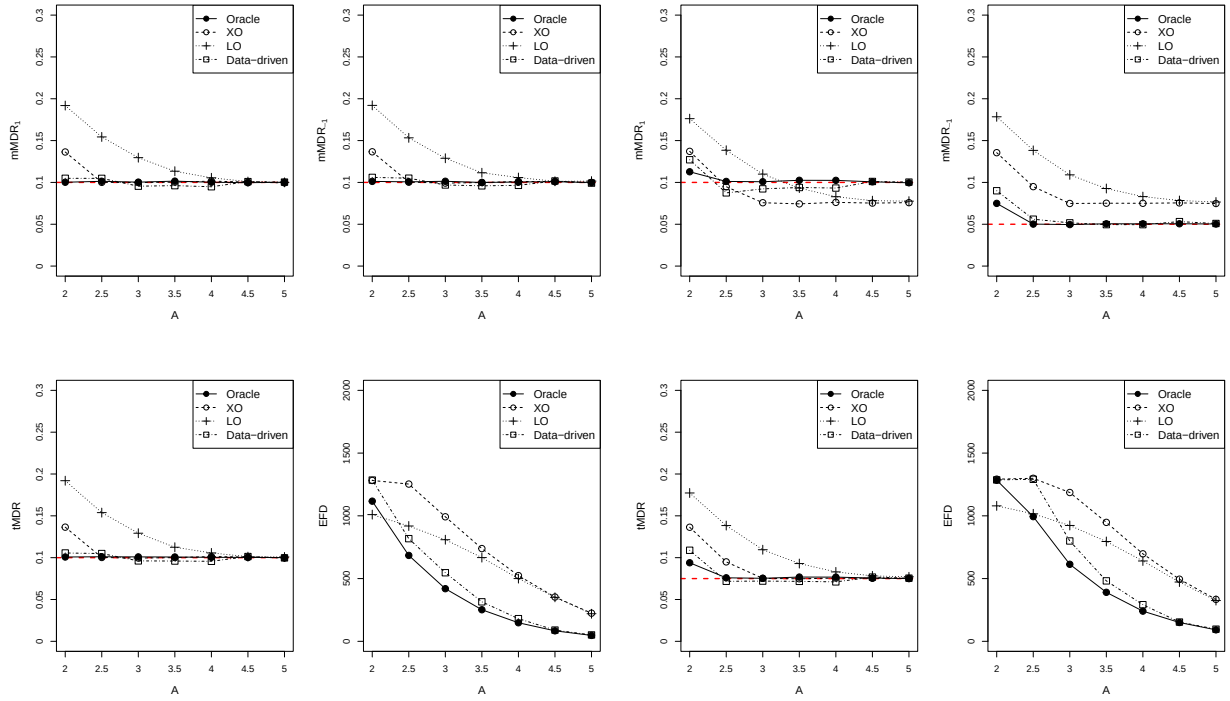
(a) $\alpha_1 = \alpha_{-1} = 0.1$ (b) $\alpha_1 = 0.1,\ \alpha_{-1} = 0.05$

Figure 2: Numerical comparison with LO and XO in the case where θ_i's are moderately dependent and the shift size varies from 1.05 to 2.55.

Next, consider the following transition matrix in the same setup.

$$\mathcal{T} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.15 & 0.7 & 0.15 \\ 0.15 & 0.15 & 0.7 \end{pmatrix}.$$

With the above transition probabilities, the OC streams occur in clusters and clumps as θ_i's are highly correlated. The results are presented in Figure 3. It can be seen that the proposed procedures outperform the other methods in terms of both controlling $\text{mMDR}_{\pm 1}$ at the given levels and achieving smaller EFD.



(a) $\alpha_1 = \alpha_{-1} = 0.1$ (b) $\alpha_1 = 0.1,\ \alpha_{-1} = 0.05$

Figure 3: Numerical comparison with LO and XO in the case where θ_i's are strongly dependent and the shift size varies from 1.05 to 2.55.

Additional numerical studies are provided in the supplementary materials, which include the impact of the number of OC observations n and the null proportion on the performance of our procedures and asymmetric transition matrix and distribution scenarios.

3.2 Normal Cases with Non-Diagonal Σ

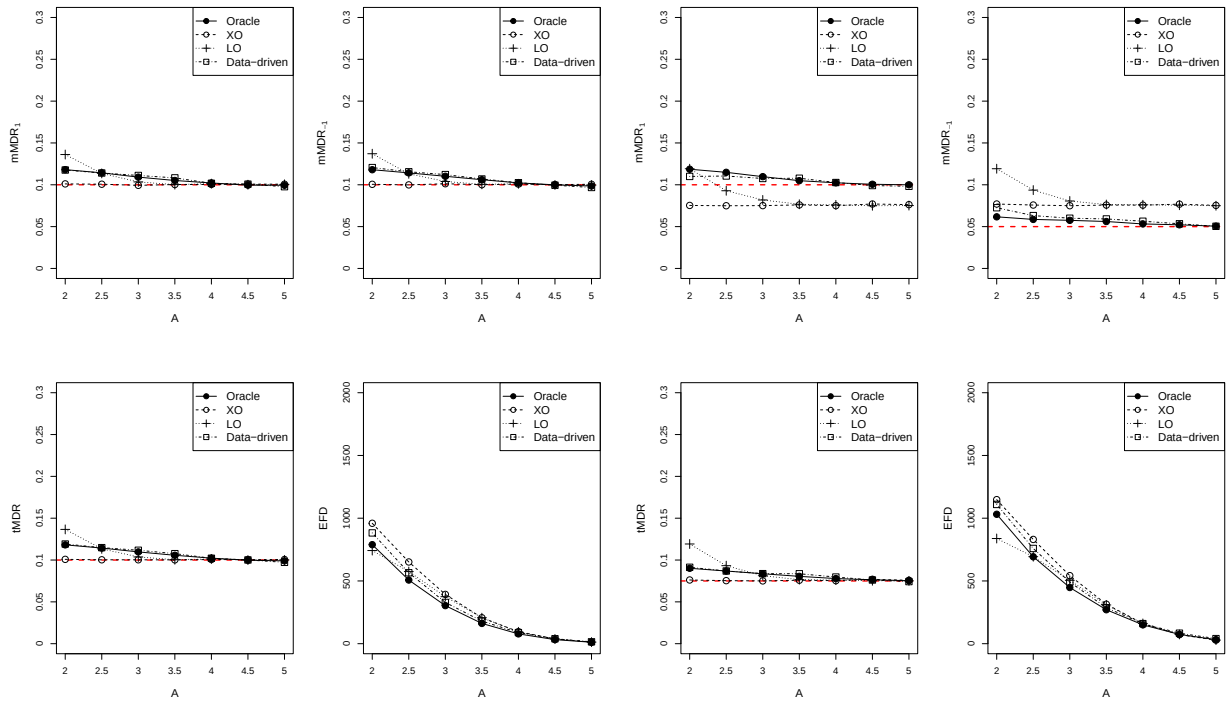
In this subsection, we assess the performance of our procedures in cases where $\boldsymbol{\Sigma}$ is non-diagonal, i.e., X_i's are conditionally correlated given $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$. Consider the following autoregressive covariance structure

$$\boldsymbol{\Sigma} = (\sigma_{i_1,i_2})_{m\times m} = \rho^{|i_1-i_2|},$$

where $\rho = 0.5$. Let $n = 5$, $\mathcal{T}$ be defined as in Figure 2, and $h_1(\mu_i) = h_2(-\mu_i) = \text{Gamma}(A, 0.05, 0.5)$ where A varies from 2 to 5. The comparison with LO and XO is shown in Figure 4, where both $\alpha_{\pm 1} = 0.1$ and $\alpha_1 = 2\alpha_{-1} = 0.1$ are considered. It can be seen from the figure that (i) XO controls $\text{mMDR}_{\pm 1}$ very well when $\alpha_1 = \alpha_{-1}$ but fails to do so when $\alpha_{\pm 1}$ are different, (ii) our procedures are able to achieve the desired levels of $\text{mMDR}_{\pm 1}$ in both settings provided that the shift size is not too small, and (iii) the proposed procedures have similar or smaller EFD in comparison with XO and LO.

3.3 Non-Normal Cases

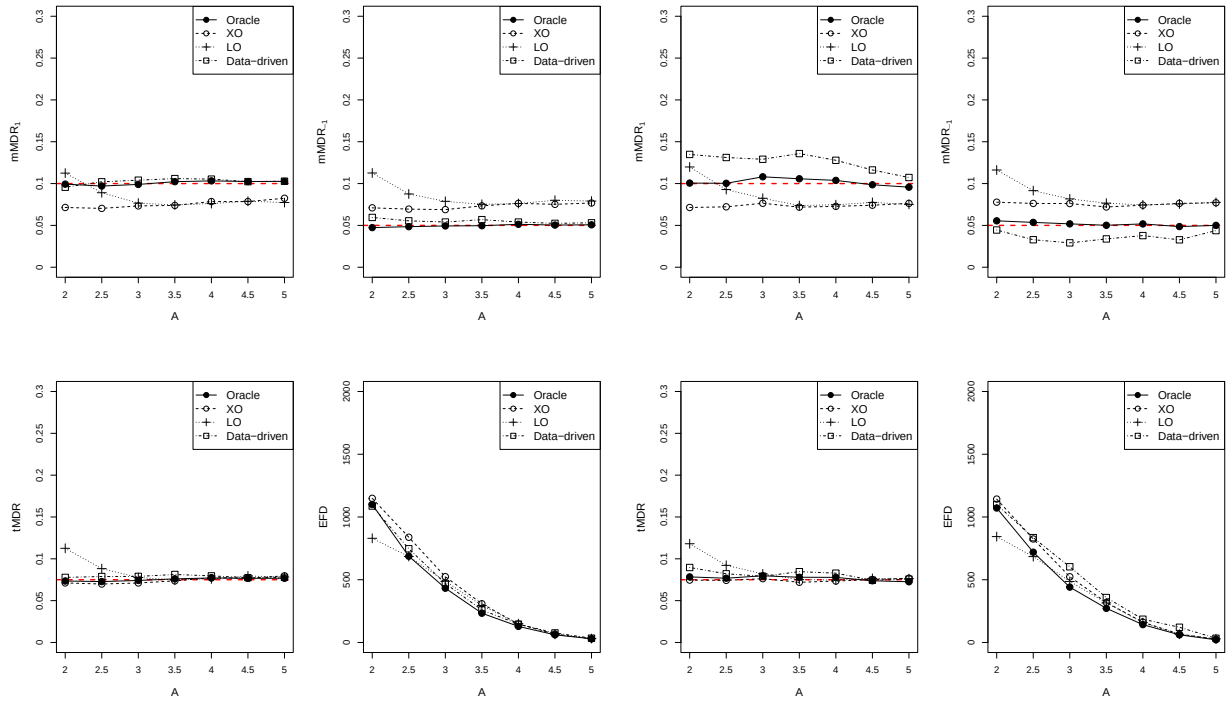
Our assumption that $\boldsymbol{X} = \boldsymbol{X}^{\texttt{OC}}/n$ follows a normal distribution is mainly justified by the central limit theorem as $\boldsymbol{X}$ is an average of the OC observations. In practice, however, this normality assumption might be violated. In this subsection, we examine our procedure in cases where the observations are non-normal. Specifically, we generate $\boldsymbol{X}_j^{\texttt{OC}}$ by (i) $\boldsymbol{X}_j^{\texttt{OC}} = \boldsymbol{X}_j'/\sqrt{5/3} + \boldsymbol{\mu}$ where $\boldsymbol{X}_j' \sim t_m(5)$, and (ii) $\boldsymbol{X}_j^{\texttt{OC}} = (\boldsymbol{X}_j' - 3)/\sqrt{3} + \boldsymbol{\mu}$ where $\boldsymbol{X}_j' \sim \text{Gamma}_m(3, 0, 1)$. Here $t_m(5)$ and Gamma_m denote the m-dimensional t distribution with 5 degrees of freedom and m-dimensional gamma distribution respectively. α_1 and α_{-1} are chosen to be 0.1 and 0.05, respectively. Our results are shown in Figure 5 where n, $\mathcal{T}$ and $\{h_1(\cdot), h_2(\cdot)\}$ are the same as those in Figure 4. It can be seen that both XO and LO fail to achieve the pre-specified $\text{mMDR}_{\pm 1}$ whereas the oracle procedure has controlled them reasonably well. The data-driven procedure performs similarly to the oracle version



(a) $\alpha_1 = \alpha_{-1} = 0.1$ (b) $\alpha_1 = 0.1,\ \alpha_{-1} = 0.05$

Figure 4: Numerical comparison with LO and XO in the case where $n = 5$, θ_i's are moderately correlated, $h_1(\mu_i) = h_2(-\mu_i) = \text{Gamma}(A, 0.05, 0.5)$, and X_i's are conditionally correlated.

in the case of t distribution but does not control $\mathrm{mMDR}_{\pm 1}$ well in the case of gamma distribution, because $\{f(\cdot|\mu_i, \theta_i)\}$ are more severely misspecified (and thus poorly estimated) in the latter scenario.



(a) $\boldsymbol{X}_j^{\mathtt{OC}}$ follows a t distribution.

(b) $\boldsymbol{X}_j^{\mathtt{OC}}$ follows a gamma distribution.

Figure 5: Numerical comparison with LO and XO in the case where $n = 5$, θ_i's are moderately correlated, $h_1(\mu_i) = h_2(-\mu_i) = \mathrm{Gamma}(A, 0.05, 0.5)$, and X_i's are not normally distributed.

3.4 Computational Cost

In this subsection, we consider the computational cost of our oracle procedure in comparison with that of XO. In the same simulation setup as in Figure 2, we examine the average computing time taken by each procedure based on 100 replications. We also record the average number of iterations for the updating algorithm to converge. The results are

summarized in Table 2. It can be seen that the oracle procedure is slightly slower than XO due to the extra time needed for calculating $H_i^k(\boldsymbol{X})$ by backward-forward formula (12) – (13) and iteratively finding $\boldsymbol{\lambda}^*$ by (10) – (11). Table 2 shows that the time difference is rather moderate even as m reaches ten thousand, suggesting that our algorithms are quite efficient. It also shows that it usually takes only a few iterations for the updating algorithm to converge. Figure 6 visualizes the results in Table 2. It can be seen that the both method's computational time grows linearly in m, indicating that they are both of computational complexity $O(m)$.

Table 2: The average computing time (in seconds) taken by the oracle procedure and XO as m varies. The unit for m is 10^3. The row *iterations* shows the average number of iterations taken for the updating algorithm (10) – (11) to converge.

m	1	2	3	4	5	6	7	8	8	10
oracle	27.16	52.35	79.64	107.27	136.09	165.27	192.52	222.50	251.79	283.06
XO	26.85	51.34	77.26	102.12	129.08	155.62	181.89	208.39	236.28	268.45
iterations	6.16	6.98	7.50	7.90	8.22	8.47	8.62	8.66	8.92	9.07

4 Real Data Example

In this section, we apply our proposed procedure to a real dataset recorded at the assembly lines of the Bosch Group (`www.bosch.com`), a global supplier in the area of special purpose machinery. The dataset was initially used in a big data competition sponsored by the company in 2016, and since then it has been made publicly available at the repository hosted by the Fraunhofer Institute (`https://www.bigdata-ai.fraunhofer.de/s/datasets/index.html`). The dataset contains 1,183,747 observations with 968 anonymized features (i.e., $m = 968$). Each observation is labeled as *pass* or *fail* based on the manu-

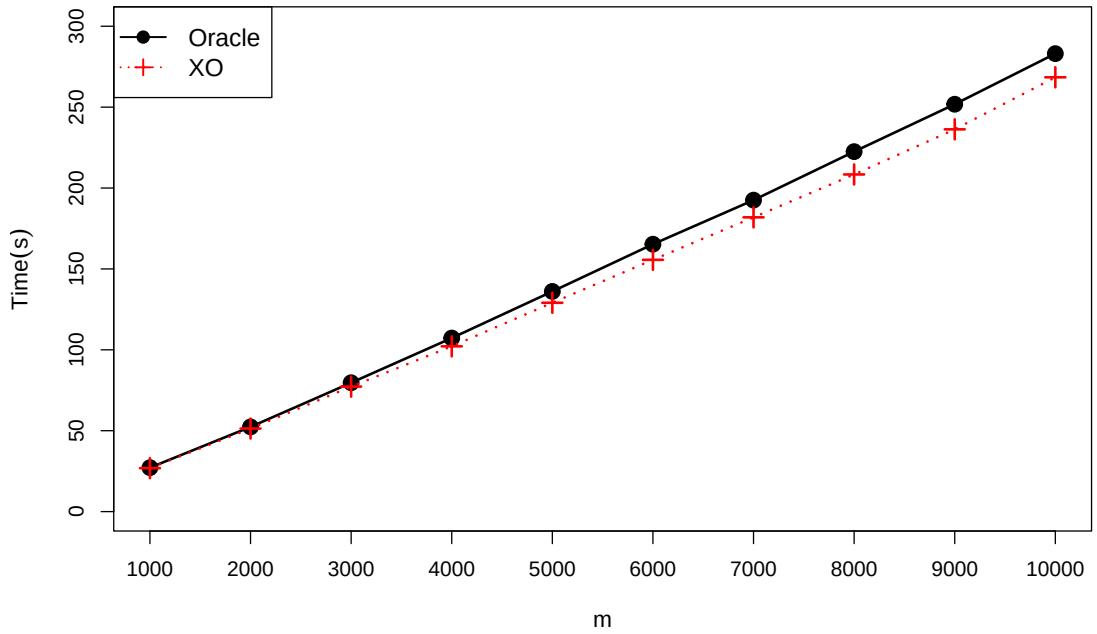


Figure 6: The graph showing the results in Table 2

facturer's internal quality and safety standards. The features are measured at a number of stations on several production lines. They are named according to the corresponding production line, the station on the line and the feature number. For example, `L3_S36_F12` is the 12th feature measured on production line 3 and station 36. Therefore, the order of these features is pre-specified.

There are 1,176,868 *pass* observations, which we regard as IC data, and 6,879 *fail* observations, which we regard as OC data (i.e., $n = 6879$). To handle the missing values in the OC observations, we adopt the following imputation method. Each missing value in the ith feature is imputed by $\widehat{F}_{1,i}^{-1}(U)$, where U is a random number from the uniform distribution on $[0, 1]$ and $\widehat{F}_{1,i}$ the empirical cumulative distribution of the ith feature computed from the OC data. Here we do not need to impute the IC data as the goal of our diagnosis is to identify the shifted features and their shift directions using the OC data only.

To ensure that the normality assumption is not severely violated, we transform the OC

data by

$$X_{ij}^{\texttt{OC}} = \Phi^{-1}\left(\widehat{F}_{0,i}\left(X'_{ij}\right)\right), \quad i = 1, 2, \ldots, m, j = 1, 2, \ldots, n,$$

where $\{X'_{ij}\}$ are the original OC observations, $\widehat{F}_{0,i}$ is the empirical cumulative distribution function of the ith feature computed from the IC data, and Φ^{-1} denotes the inverse CDF of the standard normal distribution. Figure 7 shows the between-stream correlations after the transformation. It can be seen that there are non-zero between-stream correlations, particularly among those data streams in close vicinity.

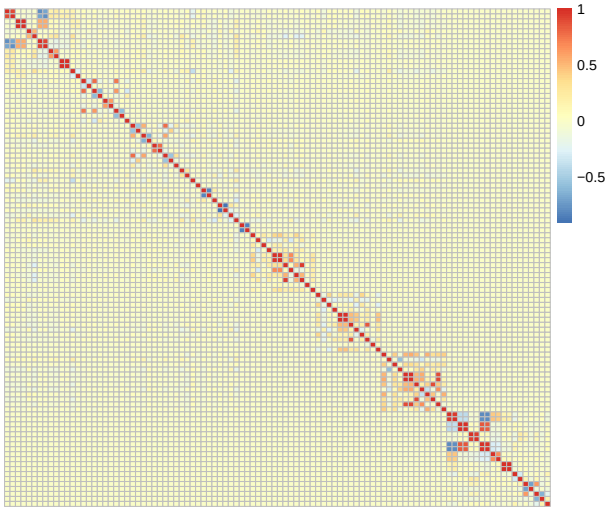


Figure 7: The between-stream sample correlation matrix of the Bosch data.

Next, we estimate our model parameters. The estimated density functions $\widehat{h}_1(\cdot)$ and $\widehat{h}_2(\cdot)$ are shown in Figure 8. It can be seen from the figure that both positive and negative shifts occurred in the process. The transition probabilities estimated by our EM algorithm are given below.

$$\widehat{\mathcal{T}} = \begin{pmatrix} 0.790 & 0.186 & 0.024 \\ 0.426 & 0.530 & 0.044 \\ 0.573 & 0.422 & 0.005 \end{pmatrix}.$$

Next, we apply the data-driven procedure with the following two sets of mMDR levels: (i) $\alpha_1 = \alpha_{-1} = 0.1$ and (ii) $2\alpha_1 = \alpha_{-1} = 0.1$. The diagnostic results are shown in Figure 9 and

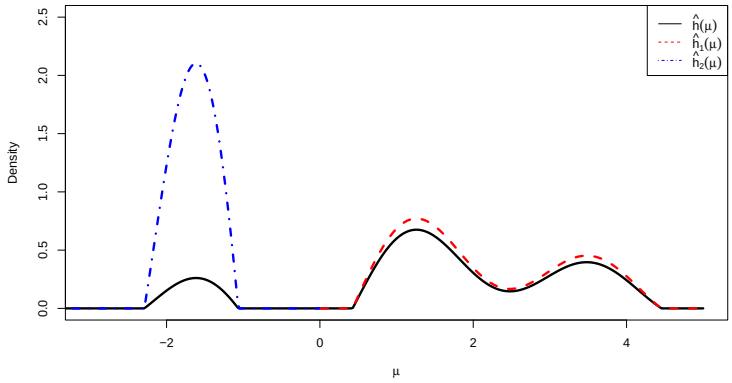


Figure 8: $\widehat{h}_1(\cdot)$ and $\widehat{h}_2(\cdot)$ for the Bosch data.

Figure 10, respectively. For comparison, the diagnostic results using XD with $\alpha = 0.1$ are shown in Figure 11 (recall that XD can not control two directional missed discovery rates separately). It is worth noting that some streams with their observed X_i values close to 0 are still determined to be OC. This is because a stream could likely be classified as OC if its neighboring streams are OC.

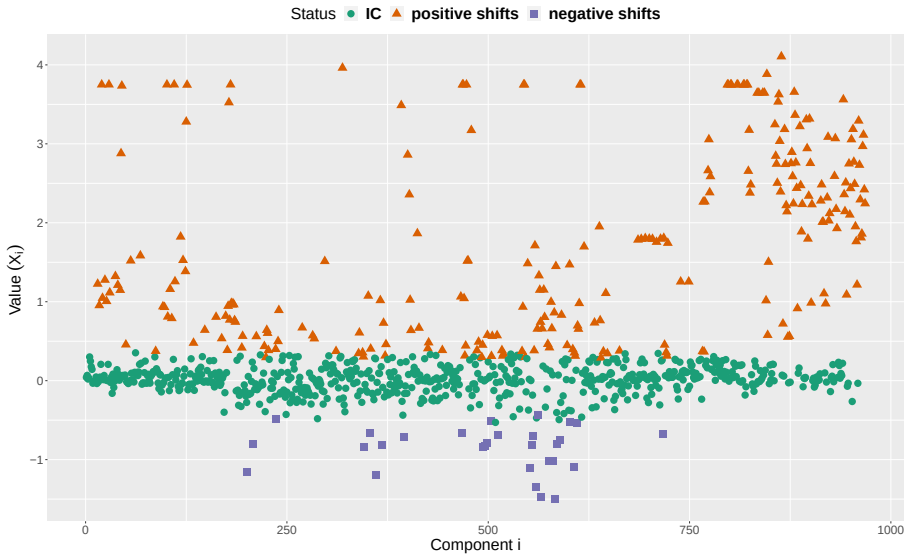


Figure 9: The Bosch data result given by the data-driven procedure with $\alpha_1 = \alpha_{-1} = 0.1$.

Due to the randomness involved in our missing value imputation, the above diagnostic results also have randomness involved. To quantify such randomness, we repeat the above

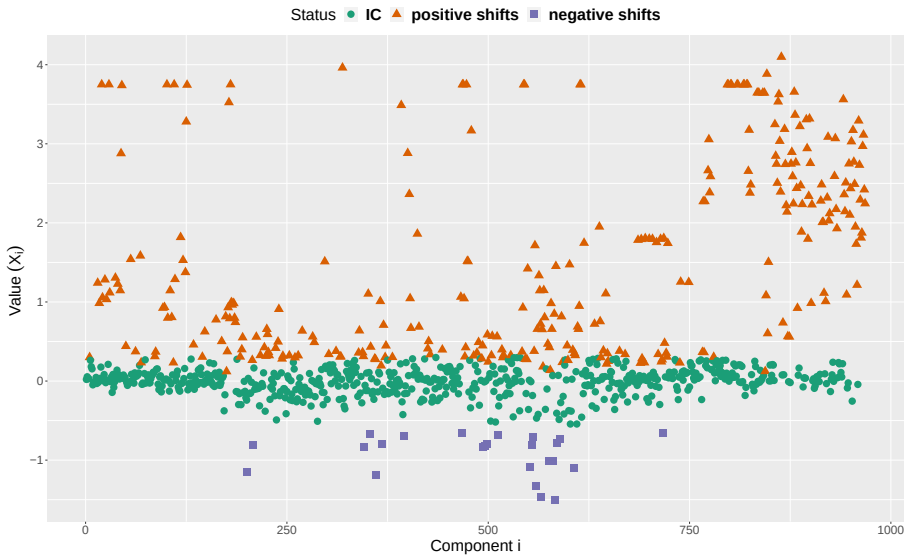


Figure 10: The Bosch data result given by the data-driven procedure with $2\alpha_1 = \alpha_{-1} = 0.1$.

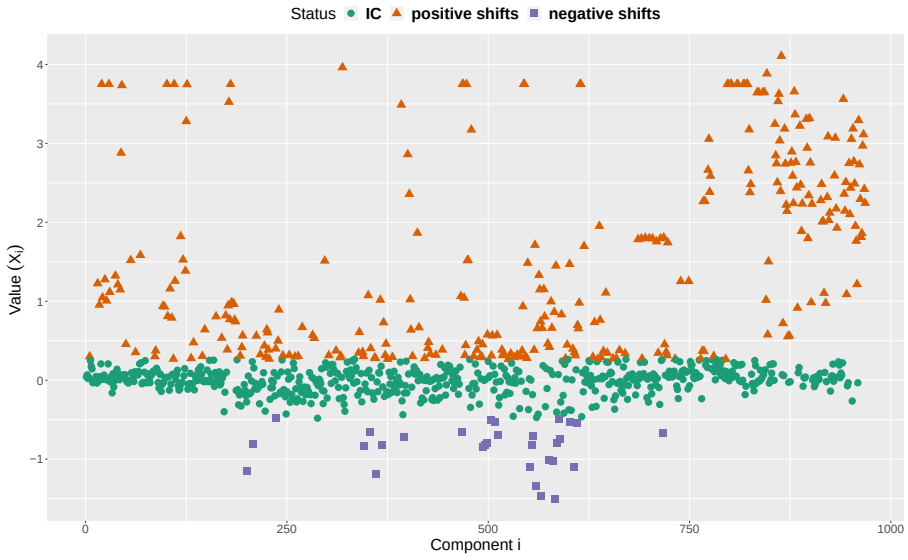


Figure 11: The Bosch data result given by XD with $\alpha = 0.1$.

analysis with 50 replicated imputations. Table 3 shows the average number of IC streams, OC streams with positive shifts and OC streams with negative shifts. The numbers in parenthesis are the corresponding standard deviations. It can be seen that the impact of the imputation randomness is relatively small. Also, more data streams are classified as having shifted in the positive direction when a smaller α_1 is used.

Table 3: Diagnostic results for the Bosch data based on 50 replicated imputations. It shows the average number of IC streams, OC streams with positive shifts and OC streams with negative shifts. The numbers in the parenthesis are the corresponding standard deviations.

Method	MDRs	$\widehat{d}_i = 0$	$\widehat{d}_i = 1$	$\widehat{d}_i = -1$
Data-driven	$\alpha_1 = \alpha_{-1} = 0.1$	676.45(6.81)	263.65(1.96)	27.90(5.10)
	$2\alpha_1 = \alpha_{-1} = 0.1$	631.252(7.95)	308.69(3.59)	28.06(4.67)
XD	$\alpha = 0.1$	634.82(9.83)	304.49(5.74)	28.69(4.53)

5 Concluding Remarks

We have proposed a fault classification procedure for high-dimensional data streams. A major feature of the proposed procedure is that it is able to simultaneously control the directional missed discovery rates at two different levels. By setting up the classification problem as a Lagrangian multiplier optimization, we have shown that our procedure is optimal in the sense that it achieves the minimum expected number of false discoveries while controlling the directional missed discovery rates at desired levels. We also suggest an iterative adjustment algorithm that converges to the optimal Lagrangian parameters. The asymptotic optimality for the data-driven version of our procedure is established as well. There are ways to further generalize our procedure. For instance, our procedure is

concerned with mean shifts only. Designing a diagnostic procedure for variance-covariance shifts is certainly an interesting direction to pursue. Additionally, in many HDDS applications, not all the data streams can be easily collected and stored due to limited computer memory. It requires future research to develop effective diagnostic procedures in such situations. Finally, the theoretical analysis of our procedure's optimality in non-normal cases is still lacking and needs to be further studied.

Acknowledgment. The authors would like to thank the Editor, an Associate Editor and two referees for many comments and suggestions which substantially improved the quality of this article.

References

Alippi, C., Ntalampiras, S., and Roveri, M. (2012). An hmm-based change detection method for intelligent embedded sensors. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Apley, D. W. and Tsung, F. (2002). The autoregressive t 2 chart for monitoring univariate autocorrelated processes. *Journal of Quality Technology*, 34(1):80–96.

Bonnett, A. H. and Soukup, G. C. (1992). Cause and analysis of stator and rotor failures in three-phase squirrel-cage induction motors. *IEEE Transactions on Industry applications*, 28(4):921–937.

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

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

Delaigle, A. and Gijbels, I. (2004). Practical bandwidth selection in deconvolution kernel density estimation. *Computational statistics & data analysis*, 45(2):249–267.

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

Diggle, P. J. and Hall, P. (1993). A fourier approach to nonparametric deconvolution of a density estimate. *Journal of the Royal statistical society: series B (Methodological)*, 55(2):523–531.

Ebrahimi, S., Ranjan, C., and Paynabar, K. (2021). Monitoring and root-cause diagnostics of high-dimensional data streams. *Journal of Quality Technology*, 54(1):20–43.

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

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272.

Fuh, C.-D. and Mei, Y. (2015). Quickest change detection and kullback-leibler divergence for two-state hidden markov models. *IEEE Transactions on Signal Processing*, 63(18):4866–4878.

Fuh, C.-D. and Tartakovsky, A. G. (2018). Asymptotic bayesian theory of quickest change detection for hidden markov models. *IEEE Transactions on Information Theory*, 65(1):511–529.

Hall, P. and Qiu, P. (2005). Discrete-transform approach to deconvolution problems. *Biometrika*, 92(1):135–148.

He, Y., Kang, Y., Tsung, F., and Xiang, D. (2023). Directional fault classification for correlated high-dimensional data streams using hidden markov models. *Journal of Quality Technology*, 55(5):535–549.

Jin, J. and 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.

Kang, Y. (2022). Statistical quality control using image intelligence: A sparse learning approach. *Naval Research Logistics (NRL)*, 69(7):996–1008.

Kulik, R. (2008). Nonparametric deconvolution problem for dependent sequences. *Electronic Journal of Statistics*, 2:722–740.

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., Xiang, D., Tsung, F., and Pu, X. (2020). A diagnostic procedure for high-dimensional data streams via missed discovery rate control. *Technometrics*, 62(1):84–100.

Li, W., Zhang, C., Tsung, F., and Mei, Y. (2021). Nonparametric monitoring of multivariate data via knn learning. *International Journal of Production Research*, 59(20):6311–6326.

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

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

Meister, A. (2009). *Deconvolution problems in nonparametric statistics*. Springer-Verlag, Berlin, Heidelberg.

Montgomery, D. C. (2020). *Introduction to statistical quality control*. John Wiley & Sons.

Qiu, P. (2014). *Introduction to statistical process control*. CRC press.

Qiu, P. (2018). Jump regression, image processing and quality control (with discussions). *Quality Engineering*, 30(1):137–153.

Qiu, P. (2020). Big data? statistical process control can help! *The American Statistician*, 74(4):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.

Randall, R. B. (2021). *Vibration-based condition monitoring: industrial, automotive and aerospace applications*. John Wiley & Sons.

Sharma, S. (2015). Food preservatives and their harmful effects. *International journal of scientific and research publications*, 5(4):1–2.

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(479):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(2):393–424.

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

Woodall, W. H. and Montgomery, D. C. (2014). Some current directions in the theory and application of statistical process monitoring. *Journal of Quality Technology*, 46(1):78–94.

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

Xiang, D., Li, W., Tsung, F., Pu, X., and Kang, Y. (2021a). Fault classification for high-dimensional data streams: A directional diagnostic framework based on multiple hypothesi s testing. *Naval Research Logistics (NRL)*, 68(7):973–987.

Xiang, D., Qiu, P., Wang, D., and Li, W. (2021b). Reliable post-signal fault diagnosis for correlated high-dimensional data streams. *Technometrics*, pages 1–12.

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.

Yi, G., Delaigle, A., and Gustafson, P. (2021). *Handbook of Measurement Error Models*. CRC Press.

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.

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(488):1586–1596.

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