Computers & Industrial Engineering 184 (2023) 109544

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

# Monitoring high-dimensional heteroscedastic processes using rank-based EWMA methods

Zezhong Wang [a,*], Rob Goedhart [b], Inez Maria Zwetsloot [a]

[a] Department of Systems Engineering, City University of Hong Kong, Hong Kong
[b] Department of Business Analytics, Amsterdam Business School, University of Amsterdam, The Netherlands

ARTICLE INFO

*Keywords:*
High-dimensional processes monitoring
Heteroscedasticity
Post signal diagnosis
Rank
Control charts
Change detection

ABSTRACT

Monitoring high-dimensional processes is a challenging task, as the underlying dependency structure among variables is often too complicated to estimate accurately. The inherent volatility of dependence, so-called heteroscedasticity, is rarely mentioned nor considered in process monitoring problems. We consider time-dependent heteroscedasticity a common cause variability and propose an integrated scheme for monitoring and diagnosis of changes in the location parameters of high-dimensional processes. Our proposed method consists of rank-based EWMA control charts which are designed to detect mean shifts in a small subset of variables. A bootstrap algorithm determines the control limits by achieving a pre-specified false alarm probability. A post-signal diagnosis strategy is executed to cluster the shifted variables and estimate a time window for the change point. Simulation results show that the proposed methodology is robust to heteroscedasticity and sensitive to small and moderate sparse mean shifts. It can efficiently identify out-of-control variables and the corresponding change points. A real-life example of monitoring online vibration data for predictive maintenance applications illustrates the proposed methodology.

## 1. Introduction

Widely used sensors and internet technology create data-rich environments. In manufacturing systems, hundreds of measurements related to production and its final products are available to evaluate their condition and quality. Statistical process monitoring (SPM) tools are prevalent for detecting persistent changes in data streams. Multivariate control charts have been proposed for monitoring multiple features simultaneously, such as the Hotelling $T^2$ control charts (Hotelling, 1947). However, they usually lose detection power as the dimension of the data increases.

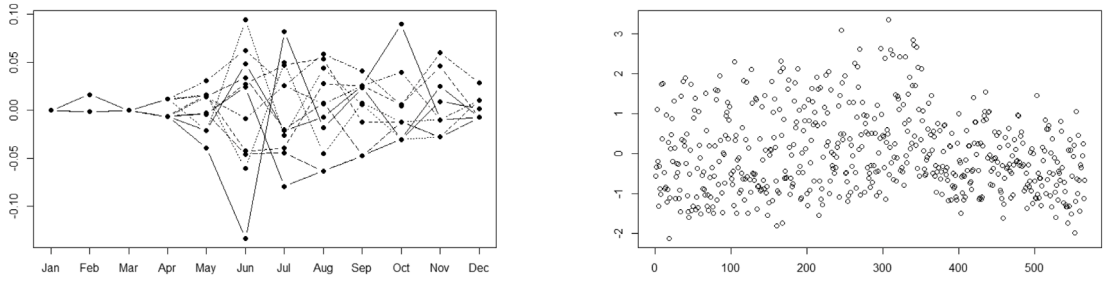
In practice, it is rare that all variables change simultaneously. It is more likely that a small subset of variables changes (Wang & Jiang, 2009). The sparse changes in mean vectors are difficult to detect using Hotelling $T^2$ methods as they are likely to be buried in the noise (Shu & Fan, 2018). One category of high-dimensional process monitoring methods is based on dimension reduction algorithms, such as principal component analysis (PCA) and variable selection (VS). De Ketelaere et al. (2015) provided a systematic review of PCA-based control charts. These techniques combine original variables linearly to get a small group of variables based on in-control (IC) data. Multivariate control charts are used to monitor these derivative variables. One limitation of PCA-based methods is that the combinations may exclude variables with important information about shifts. In addition, signal decomposition becomes increasingly difficult with the increasing number of variables (Jiang et al., 2012).

Compared with PCA-based methods, the VS-based methods select original variables directly without transformation so that the tasks of detection and diagnosis are naturally integrated and solved simultaneously (Jiang et al., 2012). Peres and Fogliatto (2018) provided an excellent overview of applying various VS algorithms in multivariate process monitoring. One of the popular VS algorithms is the least absolute shrinkage and selection operator (LASSO). It can efficiently identify suspicious out-of-control (OC) variables in the signal diagnosis procedure (Zou et al., 2011). Zou and Qiu (2009) integrated the LASSO into the multivariate exponentially weighted moving average (MEWMA) chart. Besides, Wang and Jiang (2009) used a forward variable selection (FVS) method and a generalized likelihood ratio control chart to monitor high-dimensional processes. Jiang et al. (2012) and Abdella et al. (2017) extended the same FVS method to a MEWMA chart and a multivariate cumulative sum (MCUSUM) chart, respectively.

These VS-based methods are usually based on the normality assumption, and the underlying dependency between the variables is known or can be estimated accurately. Estimating the covariance matrix is a difficult challenge in high-dimensional process monitoring

* Corresponding author.
*E-mail addresses:* zezhowang3-c@my.cityu.edu.hk (Z. Wang), r.goedhart2@uva.nl (R. Goedhart), i.m.zwetsloot@cityu.edu.hk (I.M. Zwetsloot).

https://doi.org/10.1016/j.cie.2023.109544
Received 24 November 2022; Received in revised form 3 June 2023; Accepted 14 August 2023
Available online 18 August 2023




(a) Dengue data. (b) A vibration index.

**Fig. 1.** Illustration of heteroscedasticity using different data sets.

due to the dimensionality and complicated dependence between variables (Hastie et al., 2009). Nonparametric methods are appropriate when the underlying distribution is unknown and/or difficult to verify. Rank tests, such as the Wilcoxon-type rank-sum test, are proved to be distribution-free and widely used in designing control charts (Shu & Fan, 2018). Chakraborti et al. (2001) gave an overview of rank-based univariate control charts. For more recent development of rank-based monitoring methods, see Capizzi (2015).

Rank tests are useful in detecting sparse mean shifts from multivariate and high-dimensional processes. Qiu and Hawkins (2001) proposed a MCUSUM chart based on the cross-sectional antiranks of the measurements to monitor the changes in the mean vector. The distribution of an antirank vector is known when the process is in control, so the CUSUM statistics are distribution-free. This method is valid for sparse shifts and can be adapted to high-dimensional scenarios. Li et al. (2017), Zhang et al. (2020), Zou et al. (2012) proposed different EWMA control charts for high-dimensional process monitoring based on the spatial rank test. Chen et al. (2016) used the Wilcoxon rank-sum test to design a nonparametric control chart, which can achieve satisfactory IC run-length performance for any distributions with any dimension. Mukherjee and Marozzi (2020) used the rank statistics based on the Euclidean distances of observations to design Shewhart-type nonparametric monitoring schemes.

Both the parametric and nonparametric methods for monitoring high-dimensional mean vectors usually assume that the covariance matrix is constant before and after the change point, which overlooks a potential character called heteroscedasticity in process (Hong et al., 2018). A straightforward definition of heteroscedasticity is the inequality of error variance over time (Downs & Rocke, 1979). Heteroscedasticity is rarely considered in process monitoring, though it affects the accuracy of parameter estimates and hence the performance of control charts for mean vectors. In low-dimensional processes, heteroscedastic data can be modeled by autoregressive conditional heteroscedastic (ARCH) models (Engle, 1982) or generalized autoregressive conditional heteroskedastic (GARCH) models (Bollerslev, 1986). Dispersion control charts have been developed based on the univariate or multivariate GARCH models (Bodnar, 2009; Schipper & Schmid, 2001). However, applying the GARCH model to high-dimensional scenarios is challenging because of the computational complexity (Frisén, 2008). Bai et al. (2018) used multivariate GARCH and copula to model high-dimensional time series. This method fits the time series with multivariate GARCH and then fits errors with copula's; it is computationally efficient in high-dimensional scenarios. More recently, Quevedo and Vining (2022) proposed a Shewhart control chart based on a heteroscedastic Gaussian process model for profile monitoring. To our knowledge, there is no research for detecting mean shifts in high-dimensional processes with dynamic dispersion parameters specifically. Though distribution-free methods relax the assumptions about dispersion parameters, none involve heteroscedasticity in simulations and discuss its effect.

In this paper, heteroscedasticity is assumed to be a common cause variability. When data collection procedures are affected by environmental factors or inputs, and these dynamic external impacts are difficult or impossible to remove, data will show heteroscedasticity over time. We focus on time-dependent heteroskedasticity, i.e., the covariance matrix varies with time. We define this **time-varying heteroskedasticity** as:

$$\Sigma_t = \rho_t \Sigma \tag{1}$$

Where $\rho_t$ changes with time $t$ and this change affects all variables similarly. Note that, seasonal heteroskedasticity is a part of this definition if $\rho_t$ shows a seasonal pattern.

For example, the astronomical measurements of stars are affected by the changing atmospheric effects (Tamuz et al., 2005). The speed and the stability measurements in robot experiments show heteroscedastic noise with different input parameters (Ariizumi et al., 2016). A more detailed example of heteroscedasticity as a nonremovable environmental factor is illustrated in Fig. 1(a). The mean-standardized monthly measurements of mosquito count for dengue prevention in Hong Kong are affected by seasonality variation (Wang & Zwetsloot, 2019). These residuals show heteroscedasticity over time with a 0 variance in January and a larger dispersion in summer. As the increasing mosquito counts are of interest, the seasonal pattern of variance can disturb the performance of existing monitoring methods. Fig. 1(b) plots Kurtosis of spectrum computed in Appendix A when the system operates normally from January 2021 to Aug 2021. These Kurtosis values are computed based on the vibration spectrum collected by sensors, and it is expected that vibration is affected by the system's health status as well as the workload (the number of passengers). The workload is a dynamic environmental factor and can cause heteroscedastic vibration indexes. When using the vibration index to monitor the health condition of escalators, heteroscedasticity is a vital variation that needs to be considered, as it can bury real changes. These examples show that heteroscedasticity can be an inherent characteristic in a process, which should be considered in a monitoring scheme, either removing it or via robust design. It is challenging to model and remove heteroscedasticity in high-dimensional processes. Hence we prefer a monitoring method that is robust to heteroscedasticity.

None of the aforementioned high-dimensional methods discusses or explores the effect of heteroscedasticity in the data when detecting mean shifts. In this paper, we propose a rank-based EWMA control chart for detecting sparse changes in high-dimensional mean vectors, which is robust to time-dependent heteroscedasticity. As small shifts are more likely to be buried by heteroscedastic noise and detecting small changes in high-dimensional processes is more challenging, our research focuses on monitoring small shifts. The scheme consists of a monitoring and a post-signal diagnosis stage. A data-driven algorithm is provided to obtain the control limits. After detecting a signal, the post-signal diagnosis strategy estimates the change period and the subset

of suspicious variables. Simulation and comparison results show the IC and OC performance of the proposed method. A real-data example illustrates the application of the scheme. In the case study, sensor data is collected to evaluate the health condition of essential components in escalators. Monitoring sensor data can achieve early event detection and help in preventative maintenance strategies for escalators.

The structure of this article is as follows. In Section 2, we propose our novel rank-based monitoring scheme, including the online monitoring methods and the post-signal diagnosis strategy. Section 3 shows the IC performance of the proposed methods. In Section 4, we evaluate the OC performance of the proposed methods and compare it with two existing methods. Section 5 uses the vibration case study to showcase the proposed method. Section 6 concludes the paper.

## 2. A rank-based monitoring scheme

### 2.1. Rank-based EWMA statistics

Given a series of individual $p$-dimensional observations $\boldsymbol{X}_t = (x_{t,1}, \dots, x_{t,p})', t = 1, 2, \dots$, the change-point model is

$$H_0 : \boldsymbol{X}_t \sim F_p^0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_t) \qquad \text{and} \qquad H_1 : \boldsymbol{X}_t \sim \begin{cases} F_p^0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_t), & t < \tau, \\ F_p^1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_t), & t \geq \tau. \end{cases} \tag{2}$$

Where $F_p^0$ and $F_p^1$ are the IC and OC distribution functions, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are the corresponding unknown mean vectors. We assume that sparse mean shifts occur in a subgroup $S$ with $p_1$ ($p_1 \ll p$) variables in $\boldsymbol{\mu}_1$, the shifts are sustained and fixed. And $\boldsymbol{\Sigma}_t$ is the unknown IC covariance matrix, which changes over time. In this paper, we focus on time-dependent heteroskedasticity where the variances of all variables are affected by the same time-varying factors, and all variables have the same heteroscedastic pattern over time. Thus, for each variable $j$, its standard deviation at time $t$ is given by $\sigma_{t,j} = \rho_t \sigma_j$, where $\rho_t$ is the heteroskedasticity factor at time $t$. The change-point $\tau$, which indicates the time point of the process change, is also unknown and needs to be estimated.

The ranks for $x_{t,j}$, $j = 1, \dots, p$ among $\boldsymbol{X}_t$ are integers between 1 and $p$. Practically, the elements in $\boldsymbol{\mu}_0$ are unequal, and variables with large expectations are more likely to have higher ranks. As a result, the distribution of ranks for each variable becomes incomparable. We standardize all variables so that they have approximately identical distributions at the considered time point before computing ranks. The standard scores at time point $t$ for each variable are $z_{t,j} = \frac{x_{t,j} - \bar{x}_j}{\bar{\sigma}_j}$, where $\bar{x}_j$ and $\bar{\sigma}_j$ are the IC sample mean and standard deviation of variable $j$. The primary purpose of standardization is comparison. Note that $z_{t,j}$ have an expectation of approximately 0 over time when the process is in control. As all variables share the same time-dependent heteroscedastic pattern, the influence from $\rho_t$ is the same on estimated standard deviations $\bar{\sigma}_j$ and $z_{t,j}$. Conclusively, the ranks corresponding to standard scores are comparable and approximately follow a discrete uniform distribution from 1 to $p$. This uniform distribution is identical over all variables' ranks. When $t < \tau$, the definition and distribution of the ranks $R_{t,j}$ are equal to

$$\begin{aligned} R_{t,j} &= 1 + \sum_{i \neq j} I(z_{t,j} > z_{t,i}), \\ R_{t,j} &\sim \mathcal{U}\{1, p\}. \end{aligned} \tag{3}$$

Where $\mathcal{U}\{1, p\}$ is a discrete uniform distribution with expectation $E[R] = \frac{p+1}{2}$ and variance $Var[R] = \frac{p^2-1}{12}$.

We define the sequential EWMA statistics for each variable based on $R_{t,j}$ as

$$Y_{t,j} = (1 - \lambda) Y_{t-1,j} + \lambda R_{t,j}, \tag{4}$$

with the initial value $Y_{0,j} = \frac{p+1}{2}$ and smoothing parameter $\lambda \in (0, 1]$. Under the $H_0$ hypothesis, $Y_{t,j}$ has the expectation $E[Y_{t,j}] = \frac{p+1}{2}$ and variance $Var[Y_{t,j}] = \frac{p^2-1}{12} \frac{\lambda}{1-\lambda}(1 - (1-\lambda)^{2t})$ for $j = 1, \dots, p$. Under $H_0$, $Y_{t,j}$ can be approximated by a normal distribution (Stoumbos & Sullivan, 2002),

$$F_{Y_t} \approx N\left(\frac{p+1}{2}, \frac{p^2-1}{12} \frac{\lambda}{1-\lambda}(1 - (1-\lambda)^{2t})\right). \tag{5}$$

All $Y_{t,j}$ variables follow the same $F_{Y_t}$ distribution. After the change point $\tau$, variables in $S$ will have an expected rank $E[R_{t \geq \tau, j \in S}] > \frac{p+1}{2}$ for increasing shifts and equivalently an expected rank $E[R_{t \geq \tau, j \in S}] < \frac{p+1}{2}$ for decreasing shifts. So the corresponding OC distributions of $R_{t,j}$, $j = 1, \dots, p$ are different from the IC distribution $\mathcal{U}\{1, p\}$. Conclusively, the OC distributions of $Y_{t,j}, j = 1, \dots, p$ are different from $F_{Y_t}$ so that $Y_{t,j}$ can be used to detect mean shifts in the process.

### 2.2. Online monitoring scheme

When $t > \tau$, the expectation of $Y_{t,j}$ for an increasing variable will be $E[Y_{t \geq \tau, j \in S}] > \frac{p+1}{2}$, and a decreasing variable will have $E[Y_{t \geq \tau, j \in S}] < \frac{p+1}{2}$. For non-shifted variables, their expectations are approximately equal to $\frac{p-p_1+1}{2}$, which is close to $\frac{p+1}{2}$ when $p_1 \ll p$. At a time point $t$, the ordered EWMA statistics are $Y_{t,(1)} < Y_{t,(2)} <, \dots, < Y_{t,(p)}$. $Y_{t,(p)}$ highly relates to increasing variables and can indicate increasing shifts, symmetrically, $Y_{t,(1)}$ can be used to monitor decreasing shifts. The definition of two separate monitoring statistics for decreasing and increasing shifts is as follows

$$\begin{aligned} U_t^- &= Y_{t,(1)} = min\{Y_{t,1}, Y_{t,2}, \dots, Y_{t,p}\}, \\ U_t^+ &= Y_{t,(p)} = max\{Y_{t,1}, Y_{t,2}, \dots, Y_{t,p}\}. \end{aligned} \tag{6}$$

$U_t^-$ can monitor decreasing shifts and only requires a lower control limit. And $U_t^+$ only needs an upper control limit for monitoring increasing shifts. According to the probability theory of order statistics, the IC cumulative distribution functions (CDF) for $U_t^-$ and $U_t^+$ are

$$F_{U_t^-} = 1 - [1 - F_{Y_t}]^p \text{ and } F_{U_t^+} = [F_{Y_t}]^p. \tag{7}$$

Therefore, with a predefined percentile $\alpha$, we can compute the control limits for $U_t^-$ and $U_t^+$ by

$$\begin{aligned} LCL_{U_t^-} &= F_{Y_t}^{-1}\left(1 - (1-\alpha)^{\frac{1}{p}}\right), \\ UCL_{U_t^+} &= F_{Y_t}^{-1}\left((1-\alpha)^{\frac{1}{p}}\right). \end{aligned} \tag{8}$$

Eq. (8) is an approximation. We adapt $\alpha$ to achieve target in-control performance and determine control limits. We use *LR-EWMA* and *UR-EWMA* to indicate the control charts for $U_t^-$ and $U_t^+$, respectively. We recommend running these two charts simultaneously since the directions of changes are unknown in practice. Usually, the control limits for EWMA control charts are designed to get a target IC average run length ($ARL_0$) in simulations. However, processing high-dimensional data and computing the $ARL_0$ is time-consuming. We propose a data-driven bootstrap method to determine the control limits based on false alarm probability ($FAP$). $FAP = Pr(RL \leq t | H_0)$ is the probability that at least one false alarm is observed in the process from time 1 to $t$ (Chakraborti et al., 2008). With finite observations, computing $FAP$ is less computationally intensive than computing $ARL_0$.

Algorithm 1 introduces a bootstrap method to determine the control limits with a target $FAP$. This algorithm requires a group of IC observations as input. We will analyze the effect of sample size $N$ in Section 3. Practitioners can set their own criteria to control and stop the procedure. $\epsilon$ is the step size to change $\alpha$, and a small $\epsilon$ will result in more accurate but slower convergence. $\Delta$ determines the precision of the final results. Smaller $\Delta$ can make the empirical $FAP$ closer to the preset target value.

**Algorithm 1** Bootstrap control limits

**Input** $FAP$, $\boldsymbol{X}_{N\times p}$, $\epsilon$, and $\Delta$
**Output** $\alpha$, $\boldsymbol{LCL}_{U^-}$ and $\boldsymbol{UCL}_{U^+}$
Initialize $\alpha = 0.005$, $\widehat{FAP}_{U^-} = 0$, and $\widehat{FAP}_{U^+} = 0$;
**While** $|\widehat{FAP}_{U^-} - FAP| > \Delta$, and $|\widehat{FAP}_{U^+} - FAP| > \Delta$ **do**
Compute control limits $\boldsymbol{LCL}_{U^-}$ and $\boldsymbol{UCL}_{U^+}$ based on Eq. (8)
Generate $B$ bootstrap samples $\boldsymbol{X}^b_{t\times p}$ $(b = 1, .., B)$ from $\boldsymbol{X}_{T\times p}$, each has $t$ observations
Apply the proposed monitoring scheme to all $\boldsymbol{X}^b_{t\times p}$ samples to obtain $\boldsymbol{U}^{-,b}$, and $\boldsymbol{U}^{+,b}$
$\widehat{FAP}_{U^-} \leftarrow \frac{\sum_{b=1}^{B} I(\boldsymbol{U}^{-,b} < \boldsymbol{LCL}_{U^-})}{B}$, and $\widehat{FAP}_{U^+} \leftarrow \frac{\sum_{b=1}^{B} I(\boldsymbol{U}^{+,b} > \boldsymbol{LCL}_{U^+})}{B}$
**if** $\widehat{FAP}_{U^-} - FAP > \Delta$, and $\widehat{FAP}_{U^+} - FAP > \Delta$, **then** $\alpha \leftarrow \alpha - \epsilon$;
**else if** $\widehat{FAP}_{U^-} - FAP < -\Delta$, and $\widehat{FAP}_{U^+} - FAP < -\Delta$, then $\alpha \leftarrow \alpha + \epsilon$
**end while**

### 2.3. Post signal diagnosis strategy

Upon getting a signal, it is often a challenging task to find the corresponding root causes. We propose a two-step diagnosis strategy to identify the shifted variables and estimate the change point. Identifying shifted variables can be viewed as an unsupervised classification problem. We treat each variable as an observation and cluster them into different groups. K-means clustering is suitable for this unsupervised learning task. The ranks $R_{t,j}$ value from 1 to $p$. When changes are small, the ranks of shifted and non-shifted variables may overlap. Heteroscedasticity also aggravates the overlap problem. Therefore, we cluster EWMA statistics by the following equation:

$$argmin \sum_{i=1}^{k} \sum_{\boldsymbol{Y}_{W,j}\in S_i} \|\boldsymbol{Y}_{W,j} - \boldsymbol{c}_i\|^2, \tag{9}$$

where $W$ is the window of statistics to cluster, $\boldsymbol{c}_i$ is the center for each cluster, and $k$ is the number of clusters.

Suppose that either the *LR-EWMA* or the *UR-EWMA* chart shows a signal at time point $T$ $(T > \tau)$. Likely, several data points also shifted before $T$ as EWMA statistics have a bit of a lag. Therefore, we classify the $\boldsymbol{Y}_{W,j}$ $(j = 1, \ldots, p)$, which is the vector of EWMA statistics for variable $j$ within the time window of size $W$, covering $W$ observations from $T - W + 1$ to $T$, we call this the backward window. If the diagnosis is urgent after an alarm or costly to collect more alarms, the backward window is recommended. But it may include IC data, which affects the classification results. Hence alternatively, a forward window, from $T$ to $T + W - 1$, can improve the classification results. If it is possible to collect more signals, we recommend the forward window covering more OC statistics to achieve more accurate classification and signal diagnosis.

Selecting an appropriate $W$ value is case specific. It should not be less than 3 for the k-means algorithm constraints. One possible value for $W$ is to set it equal to the number of monotonic EWMA statistics. Using the monotonically increasing EWMA statistics before the *UR-EWMA* chart signaled as the window size. And using the number of monotonically decreasing statistics before a signal in *LR-EWMA* chart as $W$ value. This paper focuses on small shifts, and it is common for the proposed methods to have some detection delay. With large shifts, we recommend using either the expected detection delay to determine $W$, waiting for a few extra observations before diagnosis, or other diagnosis strategies.

We recommend using at least $k = 3$ clusters in the k-means algorithm, one for shifted variables and the others for non-shifted variables. When the changes are small, the difference between shifted and non-shifted EWMA statistics is also small. If $p_1 \ll p$, $\hat{S}$ may include many false-positive cases with $k = 2$ clusters. Using 3 clusters can reduce this miss-classification problem. We will use simulation to support this statement. We recommend using the EWMA statistics of the signaling variable as the center for the shifted group, and the IC mean vector for one non-shifted group. For the last group, if the signal is detected by the *LR-EWMA* chart, we use the monitoring statistics in *UR-EWMA* chart as the center. Symmetrically, taking the monitoring statistics in the *LR-EWMA* chart as the third center for classification based on the *UR-EWMA* signal. One bonus of using 3 clusters is identifying potential shifts in both directions based on the alarm in one chart.

The proposed clustering strategy will return a group of potentially shifted variables $Y_{t,j}$, $j \in \hat{S}$. If EWMA statistics for selected variables show significant deviation from the IC pattern, we treat the signal as a real one and use the following estimator, proposed by Nishina (1992), to estimate the change point for each shifted variable;

$$\begin{aligned}\hat{\tau}_j^- &= argmax_t(Y_{t,j} > \frac{1+p}{2} | j \in \hat{S}, \mu_{t>\tau,j} < \mu_{t\le\tau,j}) + 1,\\ \hat{\tau}_j^+ &= argmax_t(Y_{t,j} < \frac{1+p}{2} | j \in \hat{S}, \mu_{t>\tau,j} > \mu_{t\le\tau,j}) + 1.\end{aligned} \tag{10}$$

Nishina (1992) explored the distribution of $\hat{\tau}$. Though it is a biased estimator, it is computationally efficient and suitable for high-dimensional applications. In rare situations, $Y_{t,j}$, $j \in \hat{S}$ for variable $j$ do not cross $\frac{1+p}{2}$, hence $\hat{\tau}_j^-$ and $\hat{\tau}_j^+$ are undefined. We exclude these from the estimated change-point vector $\hat{\boldsymbol{\tau}}$ $(\hat{\tau}_{j\in\hat{S}} \neq 0)$. $\hat{\boldsymbol{\tau}}^-$ is the change point vector for decreasing variables related to the *LR-EWMA* chart. $\hat{\boldsymbol{\tau}}^+$ consists of the change points for increasing variables based on the *LR-EWMA* chart. We can estimate the change period based on at least two meaningful change points;

$$\begin{aligned}\hat{\tau}^- &\in [min(\hat{\boldsymbol{\tau}}^-), max(\hat{\boldsymbol{\tau}}^-)],\\ \hat{\tau}^+ &\in [min(\hat{\boldsymbol{\tau}}^+), max(\hat{\boldsymbol{\tau}}^+)].\end{aligned} \tag{11}$$

## 3. In-control performance evaluation

We summarize the whole procedure of applying the proposed method in Fig. 2. In this section, we explore the IC performance of our proposed method under various scenarios, including heteroscedasticity. We use a multivariate normal distribution $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ as baseline model to investigate performance with various $\Sigma$ matrices. In addition, $t_{3,p}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ distributions are used to verify the robustness of the proposed method for heavy-tailed distributed data.

In our experiments, we consider $p = 50$, 100 variables with unequal expectations in $\boldsymbol{\mu}_0$. We define three different covariance matrices, $\boldsymbol{\Sigma}^1$, $\boldsymbol{\Sigma}^2$, and $\boldsymbol{\Sigma}^3$, to learn about the effect of different correlation structures on performance, where $\boldsymbol{\Sigma}^1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}^2 = (-1)^{|i-m|}(\sigma_{i,m})_{(1\le i,m\le p)}$, and $\boldsymbol{\Sigma}^3 = (\sigma_{i,m})_{(1\le i,m\le p)}$, $\sigma_{i,m} = 0.9^{|i-m|}$. We model time-dependent heteroskedasticity by multiplying the $\boldsymbol{\Sigma}$ matrix with a constant that varies over time. Details will be discussed below. To explore the impact of varying sample size $(N)$, we set $N = 50$, 100, 200, and 500.

The target $FAP$ is set at 0.1 within $t = 100$ observations. We choose $\lambda = 0.1$, $\alpha = 0.005$, $\Delta = 0.02$, $\epsilon = 0.001$ to determine the control limits with $B = 1000$ bootstrap samples using Algorithm 1. The control limits based on Algorithm 1 highly depend on the IC samples. For each scenario, we use 20 IC samples as inputs of the algorithm to compute control limits and take the average value of $\alpha$ as the final results. All performance results in this paper are obtained from $L = 1000$ simulation runs unless indicated otherwise. Table 1 reports the $\alpha$ based on 20 input samples and the corresponding average $FAP$ for IC performance. These results can be used to evaluate the reliability and robustness of Algorithm 1 and the proposed control charts.

In baseline models, $\alpha$ varies with dimensionality $p$, which has smaller values in lower-dimensional cases and larger values in higher-dimensional cases for achieving similar $FAP$. It has similar values with different $\boldsymbol{\Sigma}$, which indicates that the control limits are insensitive to the data dependency structure. The $FAP$ values show that the proposed control charts are robust to various correlations among variables. When $p = 100$, the trend that $\alpha$ increases with $N$ is substantial, such as from 0.0064 to 0.0076 under $N_{100}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}^3)$. This trend is neglectable

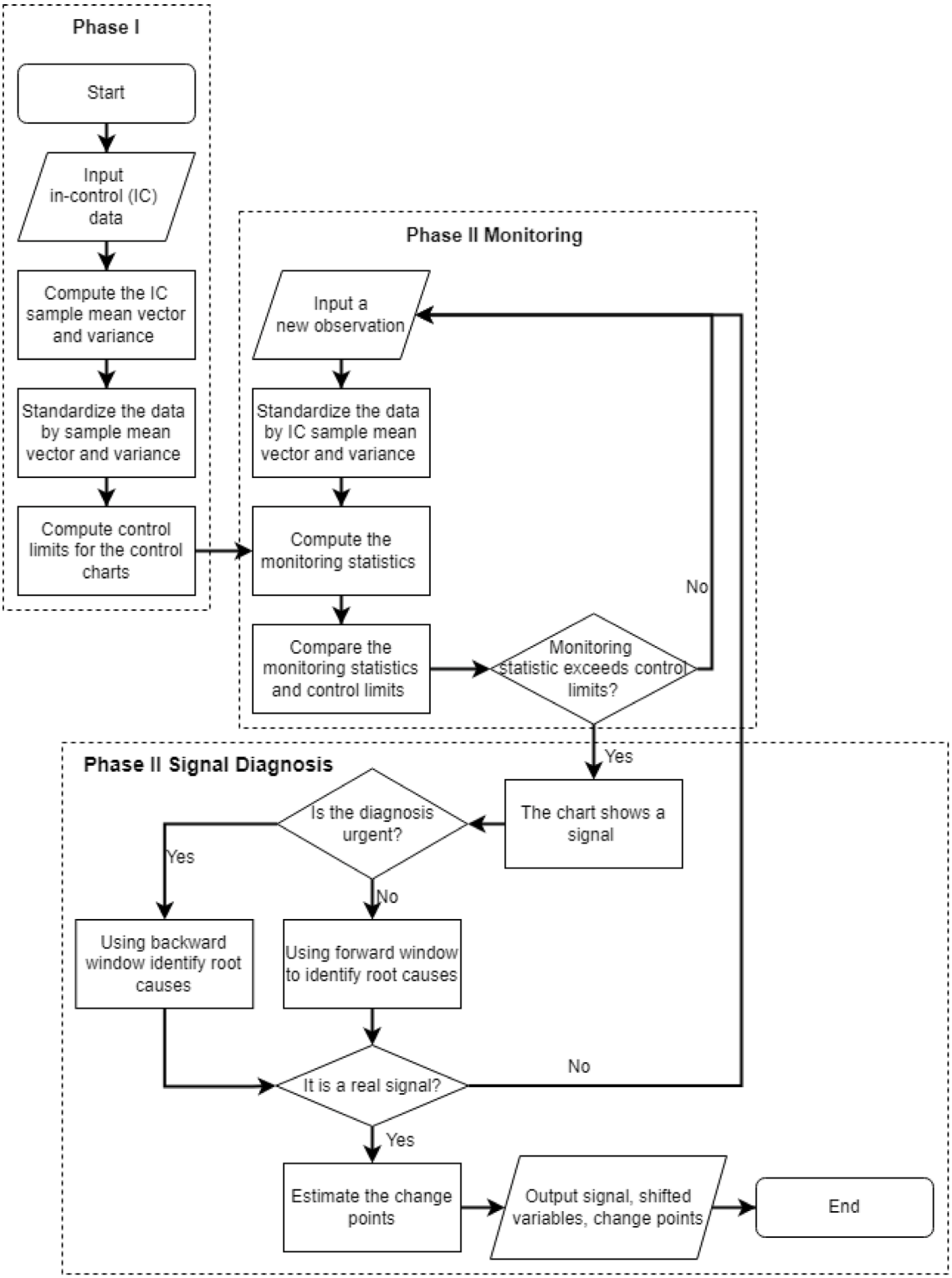



**Fig. 2.** Flow chart for the proposed monitoring method.

in low-dimensional processes because of the less noise from variables. More variables introduce more variation to parameter estimation, standardization, and results of Algorithm 1. More data can achieve more accurate parameter estimation and narrower control limits.

The effect from $p$ and $\boldsymbol{\Sigma}$ on $\alpha$ when the data are $t_{3,p}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ distributed is similar to baseline models. Heavy-tailed distributions cause smaller $\alpha$ and wider control limits than multivariate normal models. The differences are substantial with $N \leq 100$ because limited IC data result in less accurate parameter estimation and more variation. The average $FAP$ remains close to the design value of 0.1 despite the non-normal distributions and variability. Hence, our proposed methods can achieve satisfactory IC performance with non-normal distributions using the same parameters in standardization. When $N = 500$, variability caused by standardization is limited, so the $\alpha$ values converge to the baseline models. We can conclude that the proposed control charts are robust to heavy-tailed distributions.

**Table 1**
$\alpha$ and average $FAP$ values for the proposed method under normal distributions, t-distributions, heteroscedastic distribution, and varying sample size. Design $FAP = 0.1$.

| $Model$ | $p$ | $N = 50$ | | $N = 100$ | | $N = 200$ | | $N = 500$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\alpha$ | $FAP$ | $\alpha$ | $FAP$ | $\alpha$ | $FAP$ | $\alpha$ | $FAP$ |
| $N_p(\mu_0, \Sigma^1)$ | 50 | 0.0054 | 0.105 | 0.0056 | 0.099 | 0.0052 | 0.098 | 0.0055 | 0.097 |
| | 100 | 0.0058 | 0.091 | 0.0062 | 0.094 | 0.0061 | 0.091 | 0.0064 | 0.090 |
| $N_p(\mu_0, \Sigma^2)$ | 50 | 0.0054 | 0.103 | 0.0056 | 0.096 | 0.0057 | 0.095 | 0.0058 | 0.093 |
| | 100 | 0.0061 | 0.091 | 0.0067 | 0.094 | 0.0068 | 0.093 | 0.0070 | 0.091 |
| $N_p(\mu_0, \Sigma^3)$ | 50 | 0.0051 | 0.102 | 0.0055 | 0.102 | 0.0055 | 0.091 | 0.0056 | 0.090 |
| | 100 | 0.0064 | 0.093 | 0.0068 | 0.088 | 0.0072 | 0.091 | 0.0076 | 0.091 |
| $t_{3,p}(\mu_0, \Sigma^1)$ | 50 | 0.0014 | 0.097 | 0.0021 | 0.089 | 0.0034 | 0.090 | 0.0040 | 0.085 |
| | 100 | 0.0016 | 0.089 | 0.0025 | 0.092 | 0.0037 | 0.088 | 0.0047 | 0.084 |
| $t_{3,p}(\mu_0, \Sigma^2)$ | 50 | 0.0016 | 0.094 | 0.0029 | 0.092 | 0.0037 | 0.090 | 0.0045 | 0.086 |
| | 100 | 0.0020 | 0.096 | 0.0034 | 0.094 | 0.0042 | 0.090 | 0.0055 | 0.089 |
| $t_{3,p}(\mu_0, \Sigma^3)$ | 50 | 0.0017 | 0.102 | 0.0026 | 0.100 | 0.0040 | 0.091 | 0.0050 | 0.097 |
| | 100 | 0.0021 | 0.097 | 0.0033 | 0.093 | 0.0045 | 0.092 | 0.0059 | 0.087 |
| $N_p(\mu_0, \Sigma_t^1)$ | 50 | 0.0010 | 0.098 | 0.0019 | 0.089 | 0.0031 | 0.095 | 0.0038 | 0.086 |
| | 100 | 0.0010 | 0.099 | 0.0022 | 0.091 | 0.0032 | 0.089 | 0.0055 | 0.103 |
| $N_p(\mu_0, \Sigma_t^2)$ | 50 | 0.0013 | 0.096 | 0.0025 | 0.092 | 0.0034 | 0.092 | 0.0051 | 0.100 |
| | 100 | 0.0012 | 0.092 | 0.0028 | 0.091 | 0.0035 | 0.087 | 0.0057 | 0.095 |
| $N_p(\mu_0, \Sigma_t^3)$ | 50 | 0.0010 | 0.099 | 0.0022 | 0.098 | 0.0034 | 0.090 | 0.0050 | 0.099 |
| | 100 | 0.0012 | 0.094 | 0.0028 | 0.088 | 0.0038 | 0.086 | 0.0061 | 0.089 |

We use time-varying covariance matrices to model time-dependent heteroscedasticity and study the robustness. We do this by setting $\boldsymbol{\Sigma}_t = \rho_t \boldsymbol{\Sigma}$, where $\rho_t$ is set equal to $\{0.1^2, 0.2^2, \ldots, 1.8^2, 1.9^2, 1.8^2, \ldots, 0.1^2\}$ and repeat this until the end of the process. Seasonal heteroscedasticity belongs to the category of time-dependent pattern. Ideally, the average $FAP$ remains close to the design value of 0.1 despite the introduced heteroscedasticity. The effect from $p$ and $\boldsymbol{\Sigma}$ on $\alpha$ in heteroscedastic distributions is similar to baseline models. The $FAP$ values close to the design value, we can conclude that the proposed methods are robust to dependence structures under different distributions. The control limits estimated with limited IC heteroscedastic data ($N = 50$) are wider than the baseline model. $\alpha$ based on heteroscedastic samples converges to the baseline $\alpha$ with increasing $N$. One possible explanation for the difference is that heteroscedasticity affects parameter estimation and introduces more uncertainty to the standardization procedure.

We evaluate the bootstrap estimation error by considering the variability of $\alpha$ based on 20 different samples. We compute the standard deviations of $\alpha$ ($\sigma_\alpha$), which are smaller than 0.001 in all scenarios. They are relatively small compared with $\alpha$, which indicates that the results of Algorithm 1 with different IC samples are accurate. The corresponding $FAP$ values based on estimated $\alpha$ are close to 0.1, so the estimated control limits based on an individual IC sample are reliable for multivariate normal, heavy-tailed, and heteroscedastic models.

The IC performance analysis shows that the proposed methods are robust to complicated dependency structures and non-normal distributions. It can estimate control limits precisely with limited observations $N \leq p$. When the distributions are non-normal or have heteroscedastic variances over time, more data are preferred to increase the accuracy of estimated control limits. If the estimated parameters are used to standardize incoming observations, the estimation error caused by the heavy tail and heteroscedasticity is consistent and negligible. Hence the control charts can always get satisfactory IC performance with the control limits estimated from limited IC observations. An important factor for computing control limits is the IC sample size. To reduce variability, we recommend using $N = 200$ IC data points as the input of Algorithm 1 to compute reliable control limits under different distributions. In this section, we only consider symmetric distributions because our proposed methods require approximately identical distributed standard scores for each variable to compute ranks. Their performance under skewed distributions is questionable, so we do not recommend using the proposed methods for monitoring data from skewed distributions until further research proves their efficiency.

## 4. Out-of-control performance evaluation

### 4.1. Signal detection

In this section, we compare the OC performance of the proposed method with the distribution-free EWMA (*DFEWMA*) control chart proposed by Chen et al. (2016) and the interpoint distances (*IPD*) control charts proposed by Shu and Fan (2018). We choose these two control charts for comparisosn because they are distribution-free and can efficiently detect sparse shifts in location parameters in high-dimensional processes.

We determine the control limits to ensure different charts have the same $FAP$, then use two new metrics to quantify their OC performance. The first metric, the *Detection Rate (DR)*, represents the detection power. It is the proportion of runs with a signal out of all simulations, for example, $DR = \frac{\sum_{l=1}^{L} I(U_t^+ > UCL_{U_t^+} | H_1)}{L}$ for the *UR-EWMA* chart. $DR$ close to 1 indicates the method is sensitive to increasing changes. And to assess the timeliness of the monitoring scheme, we define the *Conditionally Expected Detection Delay (CED)* as the average time between the signal point and the change point. The $CED$ for the *UR-EWMA* chart is $CED = \frac{\sum_{l=1}^{L} min(arg_t(U_t^+ > UCL_{U_t^+} | H_1))}{L} - \tau$. A smaller $CED$ means faster detection.

For our comparison, we implement the *UR-EWMA* chart only, as the *UR-EWMA* and *LR-EWMA* charts are symmetric, the simulated results are also valid to the *LR-EWMA* chart. We use the same heteroscedastic models as in Section 3, but with $\boldsymbol{\mu}_0 = \mathbf{0}$. For *EWMA* type control charts, we use two smoothing parameters $\lambda = 0.1$ and 0.2 for the *UR-EWMA* chart, and $\lambda = 0.1$ for the *DFEWMA* chart. The first $p_1 = 5$ components in the location parameter change by $\delta = 0.5$ or 1 after $\tau = 100$ observations. According to the guideline in Chen et al. (2016), the reference sample for *DFEWMA* should consist of at least $m_0 = 100$ observations. As the *DFEWMA* method accumulates the information, it will use $m_0 + \tau$ observations as a baseline when the changes occur. We obtain the control limits for our method and the *IPD* charts based on $N = 200$ observations to have a similar setup. Table 2 reports the results.

Both methods are designed to achieve $FAP = 0.1$ with $t = 100$ observations. Column $\alpha\,(FAP)$ in Table 2 shows $\alpha$ and the corresponding empirical $FAP$ values for the *UR-EWMA* chart and *DFEWMA* chart. The definition of $\alpha$ in the *DFEWMA* control chart is the constant conditional false alarm rate that the chart signals with no previous alarms. Chen

**Table 2**
The $DR$ and $CED$ of the *UR-EWMA*, *DFEWMA*, and *IPD* charts under heteroscedastic distributions, with varying shift sizes. When the design $FAP = 0.1$.

| *Model* | $p$ | $\delta$ | $UR-EWMA_{\lambda=0.1}$ | | | $UR-EWMA_{\lambda=0.2}$ | | | $DFEWMA$ | | | $IPD_2$ | | | $IPD_{\text{inf}}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\alpha(FAP)$ | $DR$ | $CED$ | $\alpha(FAP)$ | $DR$ | $CED$ | $\alpha(FAP)$ | $DR$ | $CED$ | $CL(FAP)$ | $DR$ | $CED$ | $CL(FAP)$ | $DR$ | $CED$ |
| $N_p(\mathbf{0}, \Sigma_t^1)$ | 20 | 0 | 0.005(0.109) | | | 0.007(0.106) | | | 0.0003(0.106) | | | 0.069(0.121) | | | 0.075(0.095) | | |
| | | 0.5 | 1 | 14.9 | | 1 | 14.9 | | 0.798 | 31.9 | | 1 | 13.2 | | 1 | 13.5 | |
| | | 1 | 1 | 11.4 | | 1 | 11.3 | | 0.895 | 10.4 | | 1 | 9.1 | | 1 | 11.0 | |
| | 50 | 0 | 0.006(0.106) | | | 0.009(0.082) | | | 0.0002(0.094) | | | 0.062(0.089) | | | 0.080(0.120) | | |
| | | 0.5 | 1 | 13.9 | | 1 | 13.9 | | 0.541 | 44.7 | | 0.994 | 18.2 | | 1 | 14.4 | |
| | | 1 | 1 | 10.5 | | 1 | 10.6 | | 0.913 | 14.7 | | 1 | 11.0 | | 1 | 12.9 | |
| | 100 | 0 | 0.007(0.097) | | | 0.012(0.084) | | | 0.0001(0.110) | | | 0.058(0.097) | | | 0.081(0.103) | | |
| | | 0.5 | 1 | 13.9 | | 1 | 13.9 | | 0.371 | 46.0 | | 0.72 | 34.3 | | 1 | 15.7 | |
| | | 1 | 1 | 10.5 | | 1 | 10.5 | | 0.876 | 24.4 | | 1 | 13.1 | | 1 | 13.7 | |
| $N_p(\mathbf{0}, \Sigma_t^2)$ | 20 | 0 | 0.005(0.093) | | | 0.006(0.099) | | | 0.0005(0.085) | | | 0.162(0.084) | | | 0.118(0.096) | | |
| | | 0.5 | 1 | 14.8 | | 1 | 15.3 | | 0.137 | 47.4 | | 0.738 | 35.8 | | 1 | 12.9 | |
| | | 1 | 1 | 11.4 | | 1 | 11.6 | | 0.440 | 42.2 | | 1 | 11.1 | | 1 | 10.3 | |
| | 50 | 0 | 0.006(0.092) | | | 0.01(0.083) | | | 0.0005(0.092) | | | 0.120(0.076) | | | 0.115(0.114) | | |
| | | 0.5 | 1 | 13.9 | | 1 | 13.8 | | 0.111 | 48.3 | | 0.504 | 41.4 | | 1 | 14.7 | |
| | | 1 | 1 | 10.6 | | 1 | 10.6 | | 0.241 | 49.7 | | 1 | 15.4 | | 1 | 12.2 | |
| | 100 | 0 | 0.006(0.079) | | | 0.014(0.084) | | | 0.0005(0.093) | | | 0.093(0.103) | | | 0.093(0.117) | | |
| | | 0.5 | 1 | 14.1 | | 1 | 13.8 | | 0.116 | 53.5 | | 0.281 | 48.0 | | 1 | 15.2 | |
| | | 1 | 1 | 10.7 | | 1 | 10.5 | | 0.232 | 49.3 | | 0.999 | 16.0 | | 1 | 12.9 | |
| $N_p(\mathbf{0}, \Sigma_t^3)$ | 20 | 0 | 0.004(0.087) | | | 0.006(0.117) | | | 0.0005(0.082) | | | 0.163(0.079) | | | 0.114(0.091) | | |
| | | 0.5 | 1 | 14.5 | | 0.998 | 14.9 | | 0.183 | 52.8 | | 0.731 | 32.9 | | 0.985 | 14.9 | |
| | | 1 | 1 | 11.0 | | 1 | 10.9 | | 0.676 | 37.5 | | 1 | 12.2 | | 1 | 11.4 | |
| | 50 | 0 | 0.006(0.079) | | | 0.010(0.093) | | | 0.0006(0.088) | | | 0.105(0.080) | | | 0.089(0.108) | | |
| | | 0.5 | 1 | 14.9 | | 0.998 | 15.3 | | 0.123 | 44.9 | | 0.628 | 39.3 | | 0.999 | 16.2 | |
| | | 1 | 1 | 10.9 | | 1 | 10.5 | | 0.394 | 45.9 | | 1 | 13.5 | | 1 | 12.6 | |
| | 100 | 0 | 0.008(0.093) | | | 0.014(0.088) | | | 0.0005(0.092) | | | 0.081(0.097) | | | 0.077(0.108) | | |
| | | 0.5 | 1 | 15.2 | | 0.998 | 16.0 | | 0.128 | 50.7 | | 0.484 | 39.1 | | 0.997 | 17.1 | |
| | | 1 | 1 | 11.2 | | 1 | 11.1 | | 0.272 | 45.9 | | 0.999 | 16.2 | | 1 | 13.7 | |

et al. (2016) conclude that the IC run length of the *DFEWMA* chart follows the geometric distribution under a fixed covariance matrix. Therefore, for a given $\alpha$ the simulated $\widehat{FAP} \approx 1-(1-\alpha)^t$. But the equation does not hold in our comparison. For example, the $FAP = 0.1$, then the *DFEWMA* chart should have $\alpha = 0.001$ however it has much smaller $\alpha$ (see Table 2). One possible explanation is that the control limits for the *DFEWMA* control chart are based on the empirical quantile, which is directly affected by the heteroscedasticity in data. Also its IC performance is affected by correlation, $\alpha$ has larger values under $\Sigma_t^2$ and $\Sigma_t^3$. For two *IPD* control charts, we show the control limits estimated from 100,000 simulation and corresponding $FAP$ values, because they are Shewhart-type control charts with constant control limits.

The results in Table 2 show that the $DR$ of the $UR-EWMA_{\lambda=0.1}$ chart equals 1 under various dimensions and covariance structures. The $UR-EWMA_{\lambda=0.2}$ chart is affected by dependence structure as it has $DR = 0.998$ in small shifts scenarios with $\Sigma_t^3$. Hence our proposed method is sensitive to small shifts, like $\delta = 0.5$, with appropriate smoothing parameters. The *DFEWMA* method is less sensitive to detect small changes in the high-dimensional heteroscedastic process, which results in $DR < 1$. Moreover, its efficiency decreases with increased dimensionality. Correlation also significantly affects the OC performance of *DFEWMA* control chart. When $\delta = 0.5$, with complicated dependency, the $DR$ values for *DFEWMA* control chart are close to the $FAP$, which means the chart does not signal the shift at all. The $IPD_2$ control chart can efficiently monitor sparse changes in low-dimensional processes with independent variables. Its $DR$ values decrease with increasing $p$. To evaluate the effect of sparsity levels, we fix $p_1 = 5$, which means 25% variables change among 20 variables, 10% variables change among 50 variables and 5% changes among 100 variables. $IPD_2$ chart is less efficient in detecting small and sparse changes in high-dimensional processes, consistent with the conclusion in Shu and Fan (2018). The detectability of the $IPD_2$ chart for small shifts is disturbed by complex dependence structures. When $\delta = 1$, $IPD_2$ can get $DR = 1$ in most scenarios. Shu and Fan (2018) conclude that the $IPD_{\text{inf}}$ control chart is sensitive to sparse shifts; it can get $DR = 1$ in all scenarios with $\Sigma_t^1$ and $\Sigma_t^2$. Its performance is affected by complicated dependency among variables when monitoring small changes, but the results are still satisfactory.

As for the detection delay, overall, two *UR-EWMA* charts can detect a signal within 20 observations after the change point. Larger shift sizes $\delta = 1$ can reduce the detection delay. They can keep consistent performance under high-dimensional and dependent models. The difference between these two charts' performance is small. With a low detection rate under $\Sigma_t^2$ and $\Sigma_t^2$, the corresponding $CED$s of *DFEWMA* method, which are about 50, also support the deduction that the signals might be false alarms. Because we use 100 OC observations after the change point to make the $DR$ sensible. Otherwise, with infinite observations, the $DR$ will always equal 1 for each method. Besides the effect from shift size, the detection delay of $IPD_2$ chart is affected by dimensionality and dependence structures. It has larger detection delays in high-dimensional scenarios. When $\delta = 0.5$, the $IPD_2$ chart can detect a signal with more than 30 observations after the change point. Two reasons can explain this result; the first reason is that the $IPD_2$ chart is less sensitive to sparse changes. Another reason is that Shewhart-type control charts are more efficient in detecting large changes. The $CED$ values of $IPD_{\text{inf}}$ chart are smaller than 20 in all scenarios, confirming that the $IPD_{\text{inf}}$ chart efficiently detects sparse changes. But compared with the proposed $UR-EWMA$ chart, it has larger detection delays in high-dimensional scenarios with $p = 50$ and 100. Conclusively, our proposed method outperforms the *DFEWMA* method in monitoring heteroscedastic processes both in detection rate and signal speed. Though the *IPD* charts show satisfactory performance and robustness to heteroscedasticity, our proposed methods are less affected by dimensionality and dependence structures. Further, our proposed method is more sensitive to small and sparse changes.

We compute the average time of processing 200 IC observations for each method to compare their computational efficiency under various models as shown in Fig. 3. The $x$-axis is the dimensionality, and the $y$-axis is the time in seconds. Overall, the *UR-EWMA* chart and *IPD* chart are much faster than the *DFEWMA* chart, and the superiority is

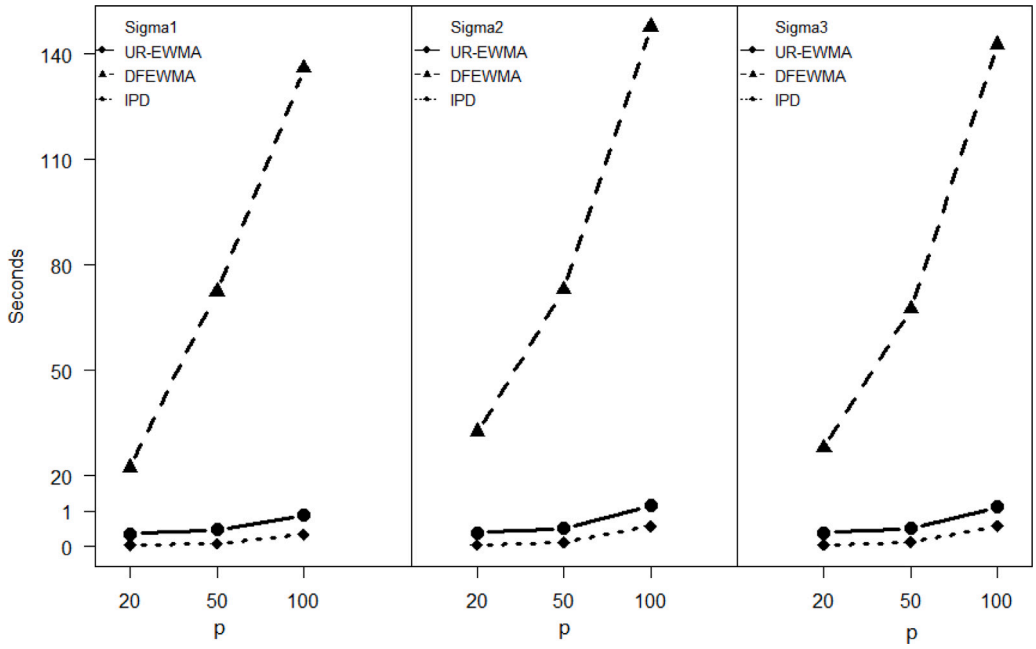



**Fig. 3.** The average processing time (seconds) of the *UR-EWMA* and the *DFEWMA* charts for monitoring 200 IC heteroscedastic data.

more prominent in high-dimensional scenarios. When $p = 100$, the *UR-EWMA* chart only costs 1 second to monitor a process but the *DFEWMA* chart needs 2 minutes. The *IPD* chart is the fastest alternative for monitoring high-dimensional processes. One convincing explanation is that individually estimating the control limits for each time point is time-consuming in the *DFEWMA* method.

The comparisons show significant advantages of our proposed method in monitoring high-dimensional heteroscedastic processes. Our proposed methods are robust to time-dependent heteroscedasticity, as they can detect small mean shifts under heteroscedastic noise with a relatively small detection delay. They are more sensitive than the Shewhart type *IPD* charts. With the superiority of $EWMA$ statistics, our methods are computationally efficient and have higher practicality in high-dimensional processes compared with the $DFEWMA$ methods. We recommend using our method for monitoring high-dimensional processes because of its robustness, sensitivity, and computational efficiency. Further, we also propose a diagnosis strategy for finding the root causes of shifts; the performances of this strategy are shown in the following section.

### 4.2. Performance of post signal diagnosis

Next, we study the performance of our signal diagnosis step. We omit a comparison with the *DFEWMA* chart and *IPD* charts from this section as they do not provide signal diagnosis strategies. After detecting a signal at time point $T$, $W$ should be smaller than $CED$ to avoid including IC statistics. Here we use $W = 5$ in clustering based on the average detection delay in Table 2. We have also studied $W = 3$ and $W = 10$, and the results are comparable, so we do not include them in this paper. We use the positive predictive rate ($PPR = \frac{TP}{TP+FP}$), and the true positive rate ($TPR = \frac{TP}{P}$) to evaluate the precision and the sensitivity of the k-means algorithm in Eq. (9). We use the average estimated period of the change points ($CPW$) to evaluate the performance of Eq. (11).

Table 3 shows the performance of the signal diagnosis strategy based on the increasing shifts with $\delta = 0.5, 1$. When $p = 20$, the backward and forward windows show $PPR \approx 1$, and both can filter the suspicious variables with limited false-positive cases. The $PPR$ values of both windows decrease with the increasing number of variables, but the forward window is less affected and can keep $PPR > 0.8$. Because the EWMA statistics in the forward window cumulates more shifted data, the deviation is more significant. Dependency has a positive effect on the $PPR$ performance. The backward window can achieve $TPR \approx 0.9$ in all scenarios, and a positive correlation can improve the $TPR$ performance. The $TPR$ values for the forward window are close to 1 in all scenarios. The k-means algorithm can identify most of the shifted variables. When $\delta = 1$, both $PPR$ and $TPR$ have larger values in all scenarios than $\delta = 0.5$, but the improvements are minor. Practitioners can choose to adjust the windows' direction and size based on the trend of the control chart.

As for the change window estimate in Eq. (11), all results can cover the real change point $\tau = 100$. The $CPW$s based on backward windows include more time points than the $CPW$s based on forward windows, especially in high-dimensional scenarios. These results may relate to the false-positive cases in the clustering step. The backward window has smaller $PPR$ values than the forward window with more variables. The performance of the forward window is less affected by dimensionality. A positive correlation can improve the estimation results since the shifts occur in 5 highly correlated variables. As shown in Table 2, the *UR-EWMA* chart has smaller $CED$ with $\delta = 1$, so the maximum estimated change points are closer to 100 compared with $\delta = 0.5$. With significant shifts, the clustering methods can identify shifted variables more accurately; hence the estimated change windows include fewer in-control points.

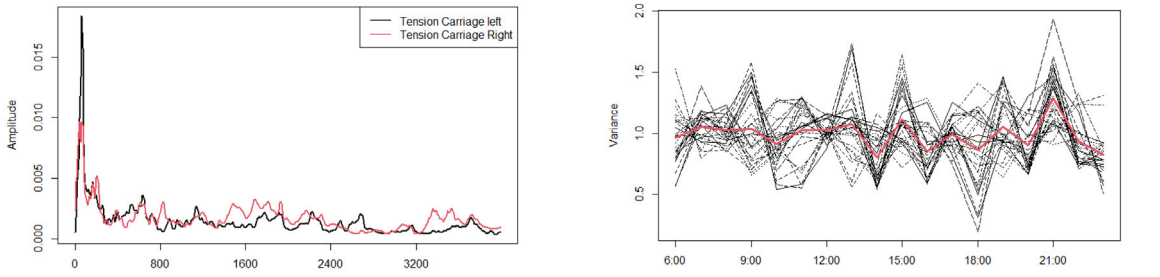
Table 3 also shows the signal diagnosis performance with $k = 2$ clusters and a forward window. All shifted variables can be detected in all scenarios, so the TPR values are close to 1. However, the PPR values are smaller than those with $k = 3$ clusters, indicating more false positive cases are included in $\hat{S}$ with $k = 2$ clusters. The misclassification problems are more significant in more sparse scenarios. Also, the estimated change windows are wider and include more time points resulting in less accurate change point estimates because of excessive false positive cases. These results support the recommendation of using 3 clusters (instead of 2) to achieve reliable diagnosis results.

## 5. Case study

We use vibration data from escalators in MTR stations to illustrate the proposed method. The Mass Transit Railway (MTR) has been operating in Hong Kong for 40 years, with more than 1000 escalators in the

**Table 3**
Post signal diagnosis for the *UR-EWMA* chart. When the design $FAP = 0.1$, $p_1 = 5$, and $\tau = 100$.

| $\delta$ | *Model* | $p$ | $(\boldsymbol{Y}_{t-W+1}, \ldots, \boldsymbol{Y}_t)_{k=3}$ | | | $(\boldsymbol{Y}_t, \ldots, \boldsymbol{Y}_{t+W-1})_{k=3}$ | | | $(\boldsymbol{Y}_t, \ldots, \boldsymbol{Y}_{t+W-1})_{k=2}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *PPR* | *TPR* | *CPW* | *PPR* | *TPR* | *CPW* | *PPR* | *TPR* | *CPW* |
| 0.5 | $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_t^1)$ | 20 | 0.93 | 0.88 | [87, 108] | 0.99 | 0.97 | [86, 109] | 0.96 | 0.98 | [86, 110] |
| | | 50 | 0.66 | 0.91 | [79, 108] | 0.93 | 0.98 | [86, 108] | 0.82 | 0.99 | [83, 109] |
| | | 100 | 0.35 | 0.93 | [71, 109] | 0.82 | 0.99 | [83, 108] | 0.53 | 1.00 | [77, 111] |
| | $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_t^2)$ | 20 | 0.95 | 0.87 | [87, 106] | 0.98 | 0.94 | [87, 107] | 0.96 | 0.97 | [88, 108] |
| | | 50 | 0.74 | 0.89 | [83, 106] | 0.94 | 0.96 | [88, 107] | 0.82 | 0.98 | [86, 108] |
| | | 100 | 0.48 | 0.91 | [77, 108] | 0.85 | 0.98 | [86, 108] | 0.60 | 0.99 | [81, 110] |
| | $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_t^3)$ | 20 | 0.97 | 0.96 | [90, 106] | 1.00 | 0.99 | [91, 106] | 0.98 | 1.00 | [91, 107] |
| | | 50 | 0.82 | 0.99 | [89, 106] | 0.96 | 0.99 | [93, 106] | 0.88 | 1.00 | [90, 106] |
| | | 100 | 0.55 | 0.99 | [82, 107] | 0.89 | 1.00 | [92, 106] | 0.67 | 0.99 | [86, 108] |
| 1 | $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_t^1)$ | 20 | 0.93 | 0.89 | [84, 104] | 0.99 | 0.98 | [85, 105] | 0.97 | 0.99 | [85, 106] |
| | | 50 | 0.67 | 0.92 | [77, 104] | 0.95 | 0.98 | [84, 105] | 0.84 | 0.99 | [82, 105] |
| | | 100 | 0.37 | 0.95 | [69, 105] | 0.86 | 0.99 | [83, 105] | 0.61 | 1.00 | [75, 106] |
| | $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_t^2)$ | 20 | 0.96 | 0.88 | [86, 103] | 0.99 | 0.95 | [86, 104] | 0.96 | 0.97 | [86, 105] |
| | | 50 | 0.76 | 0.89 | [82, 103] | 0.95 | 0.96 | [86, 104] | 0.86 | 0.98 | [84, 105] |
| | | 100 | 0.49 | 0.92 | [74, 104] | 0.88 | 0.98 | [84, 104] | 0.69 | 0.99 | [80, 106] |
| | $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_t^3)$ | 20 | 0.98 | 0.97 | [89, 103] | 0.99 | 1.00 | [89, 103] | 0.99 | 1.00 | [89, 103] |
| | | 50 | 0.82 | 0.99 | [87, 103] | 0.97 | 0.99 | [90, 102] | 0.92 | 1.00 | [90, 103] |
| | | 100 | 0.60 | 1.00 | [80, 104] | 0.94 | 1.00 | [91, 102] | 0.78 | 1.00 | [86, 104] |




(a) Vibration profiles from two sensors.

(b) Time-varying variance for health indicators.

**Fig. 4.** Vibration profiles and the heteroscedasticity.

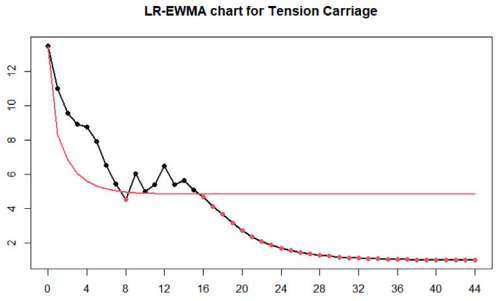
railway network. MTR conducted a project to develop a comprehensive health condition model for escalators using mathematical analysis of the related parameters. The model will be used for predictive maintenance and to support refurbishment decisions. One important escalator component is the tension carriage, which can maintain the necessary tension in the paired chains to make the escalator operate properly. To investigate the condition of the tension carriage, MTR installed two sensors to collect its vibration data. Each sensor has recorded three daily profiles from 2021-01-04 to 2021-10-11, resulting in 801 records. The original data is provided and owned by MTR.

Fig. 4(a) shows two vibration profiles collected by sensors on the same day. The $x$-axis is the frequency, and the $y$-axis is the amplitude. The spectrum can reflect the condition of the tension carriage, but they are complicated to analyze. According to the engineering comments from MTR and Shen et al. (2013), we summarize the profiles by 13 indexes (see Appendix A for more details). Consequently, 26 features are used to describe the condition of tension carriage. Since no mechanical failure is reported from January 2021 to August 2021, we take the records within that period as Phase I data. The sensor records are missing on some days, so the health indexes for these spectra are nonexistent. After removing observations with missing values, we have 566 IC data points to compute the IC parameters for each variable and control limits. Fig. 4(b) shows the time-dependent heteroscedasticity for the 26 standardized health indicators from the IC sample. The thick red line is the average variance. All health indicators seem to have a smaller variance in 13:00–15:00 and a larger variance in 20:00–22:00. The passenger load will affect the vibration data collected by sensors during normal operation. Thus, all variables are affected by the same time-varying factors and share a similar time-dependent heteroscedastic pattern. For Phase II, there are 44 observations from September to October. Each variable is standardized by the sample mean and standard deviation computed from the IC data.

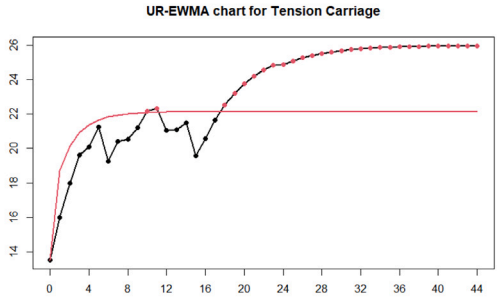
We set $\lambda = 0.2$ and $FAP = 0.1$ to obtain the control limits. Fig. 5 plots the control charts. The *LR-EWMA* chart signals at 8 and 16 until the end, and the *UR-EWMA* chart signals at 10, 11 and 18 until the end, but no reported faults are known during this period.

We diagnose the signal at 8 in *LR-EWMA* chart and the signal at 10 in *UR-EWMA* chart using forward windows. Fig. 6 plots the potential shifted variables in red. Though the k-means algorithm can identify a group of variables, there are no obvious changes in these variables. It seems that cumulative low/high ranks cause the signal. Therefore, we treat them as false alarms.

The *LR-EWMA* chart signals at 16, which is earlier than the signal at 18 in the *UR-EWMA* chart, so we use the alarm in *LR-EWMA* chart to diagnose the root causes and change points. We use a forward window with $W = 3$ to identify the shifted variables. Based on the window from 16 to 18, we can identify three suspicious decreasing variables, see Figs. 7(a) and 7(b). The estimated change window includes Sept 07. MTR reported preventive maintenance on that day. According to Figs. 7(c) and 7(d), 4 increasing variables are detected based on the window from 18 to 20 for the *UR-EWMA* chart. The change point is between Sept 07 and Sept 29. This period includes two more preventive maintenance on Sept 14 and Sept 29. On Sept 13, MTR did a half-year

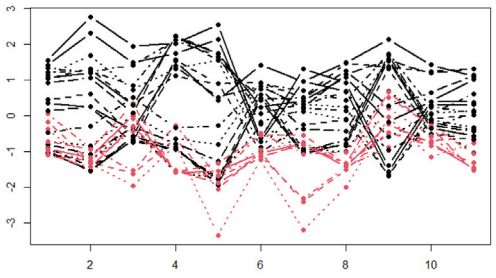



(a) LR-EWMA control chart for Phase II.

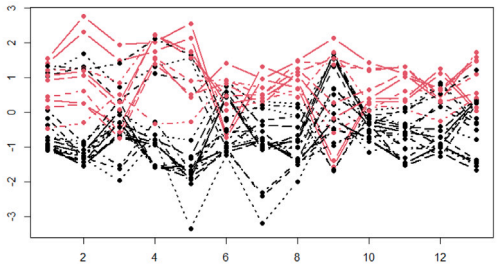



(b) UR-EWMA control chart for Phase II.

**Fig. 5.** Control charts.

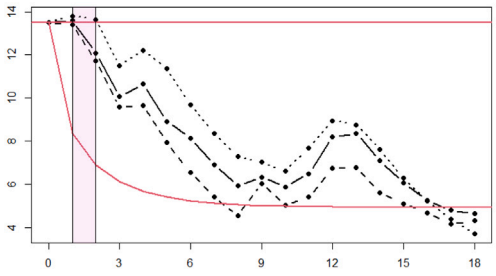


(a) Signal diagnosis LR-EWMA chart.

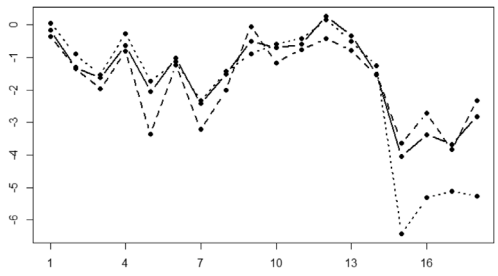


(b) Signal diagnosis UR-EWMA chart.

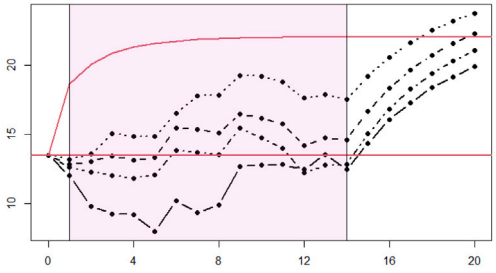
**Fig. 6.** False alarm diagnosis.



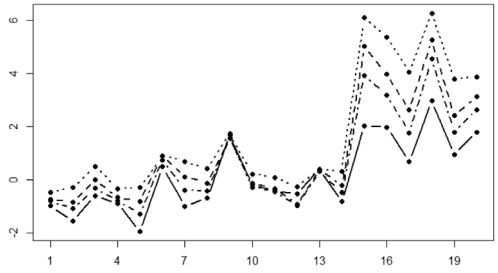
(a) $Y_{t,j}$ for three decreasing variables.



(b) Identified three decreasing variables.



(c) $Y_{t,j}$ for four increasing variables.



(d) Identified four increasing variables.

**Fig. 7.** Clustering based on LR-EWMA and UR-EWMA charts.

examination of the corresponding escalator. We can observe significant changes in the health indicator before and after that date. This case supports that the proposed monitoring scheme can detect sparse changes in high-dimensional heteroscedastic data.

## 6. Conclusion

The innovation of this paper is to consider time-dependent heteroscedasticity as a common cause variability in process monitoring of high-dimensional data. We propose a robust monitoring scheme for detecting and filtering sparse changes in high-dimensional mean vectors to cope with time-dependent heteroscedasticity. When the process is in control, the properties of the ranks for each variable are known and robust to changes in underlying process distributions. The corresponding EWMA statistics can be calculated based on ranks for each variable, and the approximate distribution is derived. We use the *UR-EWMA* and *LR-EWMA* charts to detect increasing and decreasing changes separately. A bootstrap algorithm is proposed to determine the control limits based on a predefined $FAP$ and an IC sample. A comprehensive signal diagnosis strategy is proposed to identify the shifted variables and the period of change points.

The simulation results show that the *UR-EWMA* and *LR-EWMA* charts are sensitive to small and sparse changes in the heteroscedastic process. The diagnosis strategy can cluster the suspicious variables and estimate the change points. In the comparison study, our proposed charts outperform the *DFEWMA* chart and the *IPD* charts in detecting small and sparse changes under heteroscedasticity. The data-driven control limits make our methods computationally efficient and applicable in online monitoring. The proposed scheme is easy to use in real applications, such as the illustration with sensor data.

Our approach is designed to be robust towards time-dependent heteroskedasticity. Other forms of heteroskedasticity need to be further investigated, for example, using the multivariate GARCH model and copula to generate simulation data. Both model-based and robust monitoring schemes for either mean or heteroscedasticity are worth exploring as future research directions. Furthermore, we have evaluated the performance of the proposed methods with normal and heavy-tailed distributions. Extending this study to other distributions, such as skewed distributions, is a subject of future research. The primary concentration of this paper is heteroscedasticity, and the dependence structures are relatively simple. Exploring the effect of a complicated covariance matrix on the proposed methods is another future direction.

## CRediT authorship contribution statement

**Zezhong Wang:** Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Rob Goedhart:** Methodology, Writing – review & editing, Supervision. **Inez Maria Zwetsloot:** Writing – review & editing, Supervision.

## Data availability

The data that has been used is confidential.

## Acknowledgments

The work in this paper was partially supported by a grant from the Innovation and Technology Fund, Hong Kong (ref: PRP-008-20FX), the MTR Cooperation Limited, and the Research Grants Council of the Hong Kong Special Administrative Region, China, Grant/Award Number: CityU 21215319 and City University of Hong Kong [7005567].

## Appendix A

The mathematical expression of 13 health indicators are as follows,

$$A_t = \sqrt{\frac{\sum_{i=1}^{M} x_i^2}{1.5}}$$

$$Centroid = \frac{\sum_{i=1}^{M} v_i x_i}{\sum_{i=1}^{M} x_i}$$

$$SFM = \frac{\exp(\frac{\sum_{i=1}^{M} \log(x_i)}{M})}{\bar{x}}$$

$$SH = -\frac{\sum_{i=1}^{M} (\frac{x_i}{\sum_{i=1}^{M} x_i} \times \ln(\frac{x_i}{\sum_{i=1}^{M} x_i}))}{\ln M}$$

$$Skewness = \frac{\sum_{i=1}^{M} (x_i - \bar{x})^3}{N s^3}$$

$$Kurtosis = \frac{\sum_{i=1}^{M} (x_i - \bar{x})^4}{N s^4}$$

$$Crestfactor = \frac{\max |x_i|}{\sqrt{(\sum_{i=1}^{M} x_i^2)/M}}$$

$$Clearancefactor = \frac{\max |x_i|}{(\sum_{i=1}^{M} \sqrt{|x_i|}/M)^2}$$

$$Shapefactor = \frac{\sqrt{\frac{\sum_{i=1}^{M} x_i^2}{M}}}{\frac{\sum_{i=1}^{M} |x_i|}{M}}$$

$$Impulseindicator = \frac{\max(x_i)}{\sum_{i=1}^{M} |x_i|/M}$$

$$Variance = \frac{\sum_{i=1}^{M} (x_i - \bar{x})^2}{M}$$

$$Squaredrootamplitudevalue = (\frac{\sum_{i=1}^{M} \sqrt{|x_i|}}{M})^2$$

$$Absolutemeanamplitudevalue = \frac{\sum_{i=1}^{M} |x_i|}{M}.$$

Where $x_i$ is the amplitude under frequency $v_i$. $i = 1, \ldots, M$ indicates the length of spectrum. $\bar{x} = \frac{\sum_{i=1}^{M} x_i}{M}$ is the average amplitude, and $s = \sqrt{\frac{\sum_{i=1}^{M} (x_i - \bar{x})^2}{M-1}}$ is the standard deviation of $x$.

The $A_t$ value is an index recommended by MTR. $Centroid$, $SFM$, and $SH$ are three statistics in spectral analysis. The other 9 statistical parameters are recommended by Shen et al. (2013) in vibration analysis.

## Appendix B. Supplementary materials

Codes for the *UR-EWMA* and *LR-EWMA* charts are available on GitHub https://github.com/wyfwzz/Rank-EWMA-chart.

## References

Abdella, G. M., Al-Khalifa, K. N., Kim, S., Jeong, M. K., Elsayed, E. A., & Hamouda, A. M. (2017). Variable selection-based multivariate cumulative sum control chart. *Quality and Reliability Engineering International*, *33*(3), 565–578.

Ariizumi, R., Tesch, M., Kato, K., Choset, H., & Matsuno, F. (2016). Multiobjective optimization based on expensive robotic experiments under heteroscedastic noise. *IEEE Transactions on Robotics*, *33*(2), 468–483.

Bai, Y., Dang, Y., Park, C., & Lee, T. (2018). A rolling analysis on the prediction of value at risk with multivariate GARCH and copula. *Communications for Statistical Applications and Methods*, *25*(6), 605–618.

Bodnar, O. (2009). Application of the generalized likelihood ratio test for detecting changes in the mean of multivariate GARCH processes. *Communications in Statistics. Simulation and Computation*, *38*(5), 919–938.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.

Capizzi, G. (2015). Recent advances in process monitoring: Nonparametric and variable-selection methods for phase I and phase II. *Quality Engineering*, *27*(1), 44–67.

Chakraborti, S., Human, S., & Graham, M. (2008). Phase I statistical process control charts: an overview and some results. *Quality Engineering*, *21*(1), 52–62.

Chakraborti, S., Van der Laan, P., & Bakir, S. (2001). Nonparametric control charts: an overview and some results. *Journal of Quality Technology*, *33*(3), 304–315.

Chen, N., Zi, X., & Zou, C. (2016). A distribution-free multivariate control chart. *Technometrics*, *58*(4), 448–459.

De Ketelaere, B., Hubert, M., & Schmitt, E. (2015). Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data. *Journal of Quality Technology*, *47*(4), 318–335.

Downs, G. W., & Rocke, D. M. (1979). Interpreting heteroscedasticity. *American Journal of Political Science*, 816–828.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 987–1007.

Frisén, M. (2008). *Financial surveillance, vol. 71*. Wiley Online Library.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hong, D., Balzano, L., & Fessler, J. A. (2018). Asymptotic performance of PCA for high-dimensional heteroscedastic data. *Journal of Multivariate Analysis*, *167*, 435–452.

Hotelling, H. (1947). Multivariate quality control illustrated by air testing of sample bombsights. *Techniques of Statistical Analysis*, 111–184.

Jiang, W., Wang, K., & Tsung, F. (2012). A variable-selection-based multivariate EWMA chart for process monitoring and diagnosis. *Journal of Quality Technology*, *44*(3), 209–230.

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.

Mukherjee, A., & Marozzi, M. (2020). Nonparametric phase-II control charts for monitoring high-dimensional processes with unknown parameters. *Journal of Quality Technology*, 1–21.

Nishina, K. (1992). A comparison of control charts from the viewpoint of change-point estimation. *Quality and Reliability Engineering International*, *8*(6), 537–541.

Peres, F. A. P., & Fogliatto, F. S. (2018). Variable selection methods in multivariate statistical process control: A systematic literature review. *Computers & Industrial Engineering*, *115*, 603–619.

Qiu, P., & Hawkins, D. (2001). A rank-based multivariate CUSUM procedure. *Technometrics*, *43*(2), 120–132.

Quevedo, A. V., & Vining, G. G. (2022). Online monitoring of nonlinear profiles using a Gaussian process model with heteroscedasticity. *Quality Engineering*, *34*(1), 58–74.

Schipper, S., & Schmid, W. (2001). Control charts for GARCH processes. *Nonlinear Analysis. Theory, Methods & Applications*, *47*(3), 2049–2060.

Shen, C., Wang, D., Kong, F., & Peter, W. T. (2013). Fault diagnosis of rotating machinery based on the statistical parameters of wavelet packet paving and a generic support vector regressive classifier. *Measurement*, *46*(4), 1551–1564.

Shu, L., & Fan, J. (2018). A distribution-free control chart for monitoring high-dimensional processes based on interpoint distances. *Naval Research Logistics*, *65*(4), 317–330.

Stoumbos, Z. G., & Sullivan, J. H. (2002). Robustness to non-normality of the multivariate EWMA control chart. *Journal of Quality Technology*, *34*(3), 260–276.

Tamuz, O., Mazeh, T., & Zucker, S. (2005). Correcting systematic effects in a large set of photometric light curves. *Monthly Notices of the Royal Astronomical Society*, *356*(4), 1466–1470.

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

Wang, Z., & Zwetsloot, I. M. (2019). Exploring the usefulness of functional data analysis for health surveillance. In *International workshop on intelligent statistical quality control* (pp. 247–264). Springer.

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

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., Wang, Z., & Tsung, F. (2012). A spatial rank-based multivariate EWMA control chart. *Naval Research Logistics*, *59*(2), 91–110.