Robust Monitoring of Multivariate Processes With Short-Ranged Serial Data Correlation

Xiulin Xie and Peihua Qiu
Department of Biostatistics, University of Florida
2004 Mowry Road, Gainesville, FL 32610

Abstract

Control charts are commonly used in practice for detecting distributional shifts of sequential processes. Traditional statistical process control (SPC) charts are based on the assumptions that process observations are independent and identically distributed and follow a parametric distribution when the process is in-control (IC). In practice, these assumptions are rarely valid, and it has been well demonstrated that these traditional control charts are unreliable to use when their model assumptions are invalid. To overcome this limitation, nonparametric SPC has become an active research area, and some nonparametric control charts have been developed. But, most existing nonparametric control charts are based on data ordering and/or data categorization of the original process observations, which would result in information loss in the observed data and consequently reduce the effectiveness of the related control charts. In this paper, we suggest a new multivariate online monitoring scheme, in which process observations are first sequentially decorrelated, the decorrelated data of each quality variable are then transformed using their estimated IC distribution so that the IC distribution of the transformed data would be roughly $N(0, 1)$, and finally the conventional multivariate exponentially weighted moving average (MEWMA) chart is applied to the transformed data of all quality variables for online process monitoring. This chart is self-starting in the sense that estimates of all related IC quantities are updated recursively over time. It can well accommodate stationary short-range serial data correlation, and its design is relatively simple since its control limit can be determined in advance by a Monte Carlo simulation. Because information loss due to data ordering and/or data categorization is avoided in this approach, numerical studies show that it is reliable to use and effective for process monitoring in various cases considered.

Key Words: Data decorrelation; Normalization; Recursive computation; Self-starting charts; Sequential learning; Transformation.

1 Introduction

Statistical process control (SPC) provides a major tool for online monitoring of sequential processes in the manufacturing industry, environmental monitoring, disease surveillance, and many other applications (Hawkins and Olwell 1998, Montgomery 2012, Qiu 2014). Traditional SPC charts are based on the assumptions that process observations at different time points are independent and identically distributed (i.i.d.) with a parametric distribution (e.g., normal) when the process is in-control (IC). In practice, however, these assumptions are rarely valid. This paper aims to develop a new charting scheme for online monitoring of multivariate processes in cases when process observations are serially correlated and their distribution cannot be described in advance by a parametric form.

In the SPC literature, many control charts have been developed, which can be classified roughly into the following four categories: Shewhart, cumulative sum (CUSUM), exponentially weighted moving average (EWMA), and change-point detection (CPD) charts (cf., Hawkins et al. 2003, Page 1954, Roberts 1959, Shewhart 1931). Early control charts are designed mainly for cases when the observed IC data are i.i.d. and normally distributed. In the literature, it has been well demonstrated that these conventional charts would be unreliable to use in cases when their model assumptions are invalid (e.g., Apley and Lee 2008, Chakraborti and Graham 2019, Qiu 2014). So, some recent SPC research has considered cases when some of these assumptions are violated. For instance, there have been much discussion on process monitoring of serially correlated data, and various control charts have been developed in cases when serial data correlation can be described by some parametric time series models (e.g., Capizzi and Masarotto 2008, Lee and Apley 2011) or is assumed to be stationary and short-ranged (Qiu et al. 2020, Xue and Qiu 2021). In cases when IC process distribution cannot be described properly by a parametric form, many nonparametric control charts have been proposed (e.g., Chakraborti and Graham 2019, Qiu 2018). Some of them are based on data ordering/ranking (e.g., Li et al. 2013, Qiu and Hawkins 2001, Zou et al. 2012), while the others are based on data categorization (e.g., Li 2021, Qiu 2008). However, both data ordering and data categorization could result in information loss in the observed data, which would negatively affect the effectiveness of the nonparametric control charts for online process monitoring. Therefore, much future research is needed to develop new control charts that are both reliable and effective.

In this paper, we propose a new charting scheme for online monitoring of multivariate processes in cases when process observations are serially correlated and IC process distribution cannot be described properly by a parametric form. The new method tries to avoid data ordering and/or data categorization. Instead, it tries to transform the original process observations so that a conventional control chart is (approximately) appropriate to use. It consists of the following several major components. First, process observations are sequentially decorrelated. Second, the IC distribution of the decorrelated data of each quality variable is first estimated and then the decorrelated data are transformed using their estimated IC distribution so that the IC distribution of the transformed data would be roughly N(0,1). Third, the conventional multivariate exponentially weighted moving average (MEWMA) chart is applied to the transformed data of all quality variables for online process monitoring. Fourth, this proposed chart is self-starting (Hawkins 1987) in the sense that estimates of all related IC quantities are updated recursively over time. Fifth, it can well accommodate stationary short-range serial data correlation, and its control limit can be determined in advance by a Monte Carlo simulation. Numerical studies show that this new charting scheme is reliable to use and effective for process monitoring in various cases considered, in comparison with some representative alternative approaches.

The remainder of the paper is organized as follows. In Section 2, the proposed new method will be described in detail. Some simulation studies are presented in Section 3 to evaluate its numerical performance. A real-data example to demonstrate its application is discussed in Section 4. Finally, some remarks conclude the article in Section 5.

2 Proposed Method

Assume that $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ is a vector of $p \geq 1$ numerical quality characteristics to monitor about a sequential process, and its observation at time n is $\mathbf{X}_n = (X_{n1}, X_{n2}, \ldots, X_{np})'$. To online monitor the sequential process $\{\mathbf{X}_n, n \geq 1\}$, an initial IC dataset $\mathcal{X}_{IC} = \{\mathbf{X}_{-m_0+1}, \mathbf{X}_{-m_0+2}, \ldots, \mathbf{X}_0\}$ of size m_0 is assumed to be available in advance. Another assumption needed by our proposed method is that the IC serial data correlation in the observed data is stationary and short-ranged. Namely, it is assumed that the covariance matrix $\boldsymbol{\gamma}(s) = \mathrm{Cov}(\mathbf{X}_i, \mathbf{X}_{i+s})$, for any i and s, depends on s only, and two process observations become uncorrelated if their observation times are more than b_{max} apart, where b_{max} denotes the time range of serial data correlation. The stationarity

assumption is reasonable in some SPC applications because it is often assumed that the IC process distribution, including the IC serial correlation, does not change over time in manufacturing applications. The short-range assumption implies that the correlation between two observations will disappear if their observation times are far away, which should be (approximately) true in many applications.

Our proposed method can be described intuitively as follows. First, a data decorrelation procedure is applied to the initial IC data, and an initial estimate of the IC distribution of the decorrelated data can be obtained. Second, at the current time point n during online process monitoring, the observation $\mathbf{X}_n$ is first standardized and decorrelated with previous observations, and then a transformation is applied to the decorrelated observation at time n such that the IC distribution of each component of the transformed observation would be close to normal. Third, a conventional MEWMA chart is applied to the transformed data to decide whether the process is IC or not at time n. If the control chart does not give a signal at time n, then all estimates of certain IC quantities used in the chart get updated after the IC data get expanded by combining the existing IC data with the observed data at time n. These major components of the proposed method are described in more details below.

2.1 Initial estimates of certain IC quantities

From the initial IC data $\mathcal{X}_{IC}$, we first calculate initial estimates of the IC mean $\boldsymbol{\mu}$ and the IC covariance matrix $\{\boldsymbol{\gamma}(s), 0 \le s \le b_{max}\}$. Because no parametric forms are imposed on the IC process distribution, maximum likelihood estimation is unavailable. As an alternative, we consider the following moment estimates:

$$\begin{aligned} \widehat{\boldsymbol{\mu}}^{(0)} &= \frac{1}{m_0} \sum_{i=-m_0+1}^{0} \mathbf{X}_i \qquad (1) \\ \widehat{\boldsymbol{\gamma}}^{(0)}(s) &= \frac{1}{m_0 - s} \sum_{i=-m_0+1}^{-s} \left(\mathbf{X}_{i+s} - \widehat{\boldsymbol{\mu}}^{(0)}\right)\left(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}^{(0)}\right)', \quad \text{for } 0 \le s \le b_{max}. \end{aligned}$$

Then, the initial IC data $\mathcal{X}_{IC}$ can be standardized and decorrelated by a data decorrelation algorithm described below. Let $\mathbf{W}_i = (\mathbf{X}'_{i-b}, \mathbf{X}'_{i-b+1}, \ldots, \mathbf{X}'_i)'$ be a long vector consisting of $\mathbf{X}_i$ and all its previous observations that it needs to be decorrelated with, and $\widehat{\mathbf{e}}_i = [(\mathbf{X}_{i-b} - \widehat{\boldsymbol{\mu}}^{(0)})', (\mathbf{X}_{i-b+1} - \widehat{\boldsymbol{\mu}}^{(0)})', \ldots, (\mathbf{X}_i - \widehat{\boldsymbol{\mu}}^{(0)})']'$ be the corresponding residuals, where $b = min(i + m_0 - 1, b_{max})$ and

$-m_0+1 \le i \le 0$. Then, an estimated IC covariance matrix of $\mathbf{W}_i$ is

$$\widehat{\mathbf{\Sigma}}_{i,i} = \begin{pmatrix} \widehat{\gamma}^{(0)}(0) & \cdots & \widehat{\gamma}^{(0)}(b) \\ \vdots & \ddots & \vdots \\ [\widehat{\gamma}^{(0)}(b)]' & \cdots & \widehat{\gamma}^{(0)}(0) \end{pmatrix} =: \begin{pmatrix} \widehat{\mathbf{\Sigma}}_{i-1,i-1} & \widehat{\mathbf{\Sigma}}_{i-1,i} \\ \widehat{\mathbf{\Sigma}}'_{i-1,i} & \widehat{\gamma}^{(0)}(0) \end{pmatrix}.$$

By the Cholesky decomposition of $\widehat{\mathbf{\Sigma}}_{i,i}$, we have $\mathbf{L}_i \widehat{\mathbf{\Sigma}}_{i,i} \mathbf{L}'_i = \mathbf{Q}_i$, where $\mathbf{L}_i = \begin{pmatrix} \mathbf{L}_{i-1} & \mathbf{0} \\ -\widehat{\mathbf{\Sigma}}'_{i-1,i} \widehat{\mathbf{\Sigma}}^{-1}_{i-1,i-1} & I_{p\times p} \end{pmatrix}$, $\mathbf{Q}_i = diag\{\widehat{\mathbf{D}}_{i-b}, \widehat{\mathbf{D}}_{i-b+1}, \ldots, \widehat{\mathbf{D}}_i\}$, and $\widehat{\mathbf{D}}_i = \widehat{\gamma}^{(0)}(0) - \widehat{\mathbf{\Sigma}}'_{i-1,i} \widehat{\mathbf{\Sigma}}^{-1}_{i-1,i-1} \widehat{\mathbf{\Sigma}}_{i-1,i}$. Then, the covariance matrix of $\mathbf{Q}_i^{-1/2} \mathbf{L}_i \widehat{\mathbf{e}}_i$ would be close to the identity matrix. So, the decorrelated and standardized IC observation at time i is defined to be

$$\mathbf{X}^*_i = \begin{cases} [\widehat{\gamma}^{(0)}(0)]^{-1/2} \left(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}^{(0)} \right), & \text{when } i = -m_0+1, \\ \widehat{\mathbf{D}}_i^{-1/2} \left[\mathbf{X}_i - \widehat{\boldsymbol{\mu}}^{(0)} - \widehat{\mathbf{\Sigma}}'_{i-1,i} \widehat{\mathbf{\Sigma}}^{-1}_{i-1,i-1} \widehat{\mathbf{e}}_{i-1} \right], & \text{when } i > -m_0+1. \end{cases}$$

In the above data decorrelation procedure, certain inverse matrices, including $\widehat{\mathbf{\Sigma}}^{-1}_{i-1,i-1}$, $\mathbf{Q}_i^{-1/2}$, $[\widehat{\gamma}^{(0)}(0)]^{-1/2}$ and $\widehat{\mathbf{D}}_i^{-1/2}$, need to be computed. In practice, these inverse matrices may not always exist, especially in cases when the IC sample size m_0 is small. To overcome this difficulty, we suggest using the matrix modification procedure proposed in Higham (1988) to modify a symmetric matrix to a positive definite matrix, which can be accomplished using the R function **nearPD()** in the package **Matrix**. For instance, when the matrix $\widehat{\gamma}^{(0)}(0)$ is singular, we can first use the above matrix modification approach to modify it to be a positive definite matrix, denoted as $\widetilde{\gamma}^{(0)}(0)$. Then, the inverse of $\widetilde{\gamma}^{(0)}(0)$ can be used to approximate the inverse of $\widehat{\gamma}^{(0)}(0)$. It should be pointed out that the case when the related inverse matrices do not exist is rare when $m_0 \ge p + b_{max} + 1$ based on our numerical experience. So, the matrix modification procedure would not be used often.

After the above data decorrelation procedure, the decorrelated and standardized observations $\{\mathbf{X}^*_i, i = -m_0+1, -m_0+2, \ldots, 0\}$ would be roughly i.i.d. with mean $\mathbf{0}$ and covariance matrix $I_{p\times p}$. Let $F_j(x)$ be the cumulative distribution function (cdf) of the jth component of the decorrelated data, for $j = 1, 2, \ldots, p$. Then, $F_j(x)$ can be estimated by the empirical cdf defined as follows:

$$\widehat{F}^{(0)}_j(x) = \frac{1}{m_0} \sum_{i=-m_0+1}^{0} I\left(X^*_{ij} \le x\right) \tag{2}$$

where X^*_{ij} denotes the jth component of $\mathbf{X}^*_i$, and $I(u)$ is the indicator function that equals 1 when u is "true" and 0 otherwise.

2.2 Self-starting online process monitoring

In this part, we discuss online monitoring of the process observations $\{\mathbf{X}_n, n \geq 1\}$. At the current time point n, we first decorrelate and standardize the observation $\mathbf{X}_n$ with all previous observations using a data decorrelation procedure similar to the one discussed in Subsection 2.1. The decorrelated and standardized observation at time n is denoted as $\mathbf{X}_n^*$. Then, we consider the following transformation for $\mathbf{X}_n^*$:

$$\mathbf{Z}_n = \left(\Phi^{-1}[\widehat{F}_1^{(n-1)}(X_{n1}^*)], \Phi^{-1}[\widehat{F}_2^{(n-1)}(X_{n2}^*)], \ldots, \Phi^{-1}[\widehat{F}_p^{(n-1)}(X_{np}^*)]\right), \tag{3}$$

where $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal cdf, and $\widehat{F}_j^{(n-1)}(x)$ is the empirical cdf of the jth component of the decorrelated data by the time point $n-1$ (cf., Expressions in (6) below). If the process under monitoring is IC at time n, then it is obvious that the distribution of $\widehat{F}_j^{(n-1)}(X_{nj}^*)$ would be close to Uniform[0,1], for $j = 1, 2, \ldots, p$, since $\widehat{F}_j^{(n-1)}(x)$ would be close to the true cdf of the jth component of the decorrelated data. Thus, the distribution of each component of $\mathbf{Z}_n$ would be close to $N(0,1)$. Furthermore, the p components of $\mathbf{X}_n^*$ have been decorrelated. Thus, the p components of $\mathbf{Z}_n$ would be asymptotically uncorrelated. Therefore, it is natural to apply the conventional MEWMA chart (cf., Lowry et al. 1992) to $\mathbf{Z}_n$ as follows:

$$\mathbf{E}_n = \lambda \mathbf{Z}_n + (1-\lambda)\mathbf{E}_{n-1}, \quad \text{for } n \geq 1, \tag{4}$$

where $\mathbf{E}_0 = \mathbf{0}$, and $\lambda \in (0,1]$ is a weighting parameter. Then, the chart gives a signal of process distributional shift when

$$\mathbf{E}_n' \widehat{\Sigma}_{\mathbf{E}_n}^{-1} \mathbf{E}_n > h, \tag{5}$$

where $\widehat{\Sigma}_{\mathbf{E}_n} = [\lambda/(2-\lambda)] I_{p\times p}$, and $h > 0$ is a control limit.

Because $\{\mathbf{Z}_n, n \geq 1\}$ can be regarded as a sequence of i.i.d. random vectors with the distribution of $N_p(\mathbf{0}, I_{p\times p})$ when the process under monitoring is IC, the control limit h in (5) can be determined in advance using a Monte Carlo simulation to achieve a given value of the IC average run length (ARL), denoted as ARL_0. More specifically, a sequence of random vectors can be generated from the distribution $N_p(\mathbf{0}, I_{p\times p})$. Then, the chart (4)-(5) with a given h can be applied to that sequence. The run length (RL) value, defined to be the number of observation times from the beginning of process monitoring to the signal time, can then be recorded. This simulation of online process monitoring can be repeated for B times, and the average of the corresponding B RL values can be used as the estimate of the ARL_0. Then, h can be searched so that a given level of ARL_0 is reached.

In this searching process, the bisection algorithm (Qiu 2014, Chapter 4) or its modifications (Capizzi and Masarotto 2016) can be used. As a side note, it should be pointed out that other conventional multivariate SPC charts, such as the multivariate Shewhart, CUSUM and CPD charts, can also be applied to the transformed data $\{\mathbf{Z}_n, n \geq 1\}$, although the MEWMA chart is used in (4)–(5).

If the chart (4)-(5) does not give a signal at time n, then the observation $\mathbf{X}_n$ needs to be combined with the IC dataset and the estimates of the IC cdf's $\{\widehat{F}_j(x), 1 \leq j \leq p\}$ and other IC quantities $\boldsymbol{\mu}$ and $\{\boldsymbol{\gamma}(s), 0 \leq s \leq b_{max}\}$ that are used in the construction of the chart should be updated accordingly. Because the updates are implemented at each observation time before a signal is given by the chart, efficient computation is critically important. To this end, the following formulas for recursive updates of the estimates are derived: for $1 \leq j \leq p$ and $0 \leq s \leq b_{max}$,

$$\begin{aligned} \widehat{F}_j^{(n)}(x) &= \frac{m_0+n-1}{m_0+n}\widehat{F}_j^{(n-1)}(x) + \frac{1}{m_0+n} I\left(X_{nj}^* \leq x\right), \\ \widehat{\boldsymbol{\mu}}^{(n)} &= \frac{1}{m_0+n}\mathbf{X}_n + \frac{m_0+n-1}{m_0+n}\widehat{\boldsymbol{\mu}}^{(n-1)}, \\ \widehat{\boldsymbol{\gamma}}^{(n)}(s) &= \frac{1}{m_0+n-s}\left(\mathbf{X}_n - \widehat{\boldsymbol{\mu}}^{(n)}\right)\left(\mathbf{X}_{n-s} - \widehat{\boldsymbol{\mu}}^{(n)}\right)' + \frac{m_0+n-s-1}{m_0+n-s}\widehat{\boldsymbol{\gamma}}^{(n-1)}(s). \end{aligned} \tag{6}$$

The proposed self-starting online monitoring scheme can then be summarized below.

Step 1 Initial Estimation of IC Quantities: Obtain the initial estimates $\widehat{\boldsymbol{\mu}}^{(0)}$, $\{\widehat{\boldsymbol{\gamma}}^{(0)}(s), 0 \leq s \leq b_{max}\}$ and $\{\widehat{F}_j^{(0)}(x), 1 \leq j \leq p\}$ from the initial IC data $\mathcal{X}_{IC}$, as discussed in Subsection 2.1.

Step 2 Data Decorrelation and Standardization: At the current time point n, if $n = 1$, then define the standardized observation to be

$$\mathbf{X}_1^* = [\widehat{\boldsymbol{\gamma}}^{(0)}(0)]^{-1/2}\left(\mathbf{X}_1 - \widehat{\boldsymbol{\mu}}^{(0)}\right).$$

Otherwise, the estimated covariance matrix of $(\mathbf{X}'_{n-b}, \mathbf{X}'_{n-b+1}, \ldots, \mathbf{X}'_n)'$ is defined to be

$$\widehat{\boldsymbol{\Sigma}}_{n,n} = \begin{pmatrix} \widehat{\boldsymbol{\gamma}}^{(n-1)}(0) & \cdots & \widehat{\boldsymbol{\gamma}}^{(n-1)}(b) \\ \vdots & \ddots & \vdots \\ [\widehat{\boldsymbol{\gamma}}^{(n-1)}(b)]' & \cdots & \widehat{\boldsymbol{\gamma}}^{(n-1)}(0) \end{pmatrix} =: \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{n-1,n-1} & \widehat{\boldsymbol{\Sigma}}_{n-1,n} \\ \widehat{\boldsymbol{\Sigma}}'_{n-1,n} & \widehat{\boldsymbol{\gamma}}^{(n-1)}(0) \end{pmatrix},$$

where $b = \min(n-1, b_{max})$. Then, the decorrelated and standardized observation at time n is defined to be

$$\mathbf{X}_n^* = \widehat{\mathbf{D}}_n^{-1/2}\left[-\widehat{\boldsymbol{\Sigma}}'_{n-1,n}\widehat{\boldsymbol{\Sigma}}_{n-1,n-1}^{-1}\widehat{\mathbf{e}}_{n-1} + (\mathbf{X}_n - \widehat{\boldsymbol{\mu}}^{(n-1)})\right],$$

where $\widehat{\mathbf{D}}_n = \widehat{\boldsymbol{\gamma}}^{(n-1)}(0) - \widehat{\boldsymbol{\Sigma}}'_{n-1,n}\widehat{\boldsymbol{\Sigma}}^{-1}_{n-1,n-1}\widehat{\boldsymbol{\Sigma}}_{n-1,n}$, and $\widehat{\mathbf{e}}_{n-1} = [(\mathbf{X}_{n-b} - \widehat{\boldsymbol{\mu}}^{(n-1)})', (\mathbf{X}_{n-b+1} - \widehat{\boldsymbol{\mu}}^{(n-1)})', \ldots, (\mathbf{X}_{n-1} - \widehat{\boldsymbol{\mu}}^{(n-1)})']'$.

Step 3 Decision-Making: Compute the transformed data $\{\mathbf{Z}_n, n \geq 1\}$ by (3) and then apply the MEWMA chart (4)-(5) to the transformed data. The chart gives a signal when (5) is true.

Step 4 Recursive Update of Estimates of IC Quantities: If the chart (4)-(5) does not give a signal at the current time point n, then estimates of certain IC quantities should be updated by the formulas in (6).

3 Simulation Studies

In this section, we investigate the numerical performance of the proposed chart (4)-(5), denoted as NEW, for online process monitoring. In the numerical studies, besides NEW, the following four competitive charts are also considered for comparison purpose.

- The self-starting multivariate EWMA chart suggested by Hawkins and Maboudou-Tchao (2007), denoted as SS-MEWMA. The charting statistic of SS-MEWMA is

$$\mathbf{E}_{n,ss} = \lambda_{ss}(\mathbf{X}_n - \widehat{\mu}^{(n-1)}) + (1 - \lambda_{ss})\mathbf{E}_{n-1,ss}, \text{ for } n \geq 1,$$

  where $\mathbf{E}_{0,ss} = \mathbf{0}$, and $\lambda_{ss} \in (0, 1]$ is a weighting parameter. The chart gives a signal when

$$\mathbf{E}'_{n,ss}\widehat{\Sigma}^{-1}_{\mathbf{E}_{n,ss}}\mathbf{E}_{n,ss} > h_{ss},$$

  where $\widehat{\Sigma}_{\mathbf{E}_{n,ss}} = [\lambda_{ss}/(2 - \lambda_{ss})]\widehat{\gamma}^{(n-1)}(0)$, and $h_{ss} > 0$ is a control limit. This chart assumes that the original process observations are i.i.d. and normally distributed when the process is IC. So, its control limit can be determined by Monte Carlo simulations based on the assumed IC normal distribution.

- The nonparametric multivariate EWMA chart suggested by Zou et al. (2012), denoted as SR-MEWMA. This chart is based on spatial ranks of the process observations. More specifically, the spatial sign of the observation $\mathbf{X}_n$ is defined to be $\mathbf{U}(\mathbf{X}_n) = \mathbf{X}_n\|\mathbf{X}_n\|^{-1}$ when $\mathbf{X}_n \neq 0$, and $\mathbf{0}$ otherwise, where $\|\mathbf{X}_n\|$ denotes the Euclidean length of $\mathbf{X}_n$. Then, the spatial rank

of $\mathbf{X}_n$ is defined to be $\mathbf{R}_n(\mathbf{X}_n) = \frac{1}{m_0+n-1} \sum_{i=-m_0+1}^{n-1} \mathbf{U}(\mathbf{X}_n - \mathbf{X}_i)$, and the charting statistic of SR-MEWMA is defined to be

$$\mathbf{E}_{n,sr} = \lambda_{sr}\mathbf{R}_n\left([\widehat{\gamma}^{(n-1)}(0)]^{-1/2}\mathbf{X}_n\right) + (1-\lambda_{sr})\mathbf{E}_{n-1,sr}, \text{ for } n \geq 1,$$

where $\mathbf{E}_{0,sr} = \mathbf{0}$, $\lambda_{sr} \in (0,1]$ is a weighting parameter. The chart gives a signal when

$$\mathbf{E}'_{n,sr}\widehat{\Sigma}^{-1}_{\mathbf{E}_{n,sr}}\mathbf{E}_{n,sr} > h_{sr},$$

where $h_{sr} > 0$ is a control limit, and $\widehat{\Sigma}_{\mathbf{E}_{n,sr}}$ is the moment estimate of the covariance matrix of $\mathbf{E}_{n,sr}$ that can be first estimated from the initial IC data and then updated recursively. Thus, SR-MEWMA is a self-starting nonparametric control chart. But, it assumes that process observations at different time points are independent.

- The nonparametric CPD chart suggested by Holland and Hawkins (2014), denoted as SR-CPD. This chart is constructed based on the multivariate spatial rank test. Its charting statistic is defined to be

$$\max_{1\leq k<n-c}\left\{\bar{\mathbf{r}}_n^{(k)\prime}\left[\frac{n+m_0-k}{(n+m_0)k}\widehat{\gamma}^{(n-1)}(0)\right]^{-1}\bar{\mathbf{r}}_n^{(k)}\right\},$$

where c is the pre-specified number of observations at the end of the sequence that will not be considered for a possible change point, and $\bar{\mathbf{r}}_n^{(k)} = \frac{1}{k+m_0}\sum_{i=-m_0+1}^{k}\mathbf{R}_n(\mathbf{X}_i)$. This chart assumes that process observations at different time points are independent.

- The multivariate nonparametric CUSUM chart suggested by Xue and Qiu (2021), denoted as XQ-CUSUM. The chart first decorrelates the observed data, and then applies the multivariate nonparametric chart based on data categorization (cf., Qiu 2008) to the decorrelated data for online process monitoring.

Regarding the IC process distribution and the IC serial data correlation, the following four cases when $p = 3$ are considered.

Case I: Process observations $\{\mathbf{X}_n, n \geq 1\}$ are i.i.d. with the IC distribution $N_3(\mathbf{0}, I_{3\times 3})$.

Case II: Process observations $\mathbf{X}_n = (X_{n1}, X_{n2}, X_{n3})'$ are i.i.d. at different observation times. Their three components X_{n1}, X_{n2} and X_{n3} are independent with the distributions $N(0,1)$, the standardized version with mean 0 and variance 1 of the χ^2_3 distribution, and the standardized version with mean 0 and variance 1 of the t_3 distribution, respectively.

Case III: Process observations $\{\mathbf{X}_n, n \geq 1\}$ follow the Vector AR(1) model $\mathbf{X}_n = \mathbf{A}\mathbf{X}_{n-1} + \boldsymbol{\epsilon}_n$, where A is a diagonal matrix with the diagonal elements 0.3, 0.2 and 0.1, and the p components of the error term $\{\boldsymbol{\epsilon}_n\}$ are independent with the distributions $N(0,1)$, the standardized version with mean 0 and variance 1 of the χ^2_3 distribution, and the standardized version with mean 0 and variance 1 of the t_3 distribution, respectively.

Case IV: Process observations $\{\mathbf{X}_n, n \geq 1\}$ follow the Vector AR(1) model $\mathbf{X}_n = \mathbf{A}\mathbf{X}_{n-1} + \mathbf{C}^{1/2}\boldsymbol{\epsilon}_n$, where A is a diagonal matrix with the diagonal elements being 0.3, 0.2 and 0.1, $\{\boldsymbol{\epsilon}_n\}$ are generated in the same way as that in Case III, and C is

$$\begin{pmatrix} 1 & 0.2 & 0.2^2 \\ 0.2 & 1 & 0.2 \\ 0.2^2 & 0.2 & 1 \end{pmatrix}.$$

About the four cases described above, Case I is the conventional case considered in the SPC literature with i.i.d. process observations and the standard normal IC process distribution. Case II considers a scenario when the IC distributions of some quality variables are not normal. Cases III and IV consider two scenarios with stationary serial data correlation when the p quality variables are independent (Case III) or mutually associated (Case IV).

Evaluation of the IC performance: We first evaluate the IC performance of the related control charts. In the simulation study, the IC sample size m_0 can change among $\{100, 200, 300, 500\}$. The nominal ARL_0 values of all charts are fixed at 200. The weighting parameters in the charts SS-MEWMA, SR-MEWMA and NEW are fixed at 0.05, and the allowance constant in the chart XQ-CUSUM is chosen to be 0.1. In the charts XQ-CUSUM and NEW, b_{max} is chosen to be 10. The control limits of SS-MEWMA and NEW are determined by simulation as discussed earlier, and the control limits of SR-MEWMA, SR-CPD and XQ-CUSUM are computed as discussed in Zou et al. (2012), Holland and Hawkins (2014) and Xue and Qiu (2021). For each method, its actual ARL_0 value is computed as follows. First, an IC dataset of size m_0 is generated, and some IC parameters are estimated from the IC dataset. Then, each control chart is applied to a sequence of 2,000 IC process observations for online process monitoring, and the RL value is recorded. This simulation of online process monitoring is then repeated for 1,000 times, and the actual conditional ARL_0 value conditional on the IC data is computed as the average of the 1,000 RL values. The entire simulation described above, from generation of the IC dataset to computation of the conditional ARL_0 value, is then repeated for 100 times. The average of the 100 actual conditional ARL_0 values

is used as the estimated actual ARL_0 value of the related control chart, and the standard error of this estimated actual ARL_0 value can also be computed. The estimated ARL_0 values in different cases considered are shown in Table 1.

From Table 1, we can have the following conclusions. First, SS-MEWMA performs well in Case I since all its model assumptions are valid in that case. In all other cases, it does not perform well since some of its model assumptions (e.g., normality, data independence) are violated in those cases. Second, the performance of SR-MEWMA and SR-CPD is good in Cases I and II when process observations are independent at different time points. But, their actual ARL_0 values are substantially different from the nominal ARL_0 value of 200 in the other two cases when their "data independence" assumption is violated. Third, the charts XQ-CUSUM and NEW both have a quite reliable IC performance in all cases considered when $m_0 \geq 300$ since their actual ARL_0 values are within 5% of the nominal ARL_0 value in such cases. Remember that the chart XQ-CUSUM is based on nonparametric process monitoring by data categorization while the chart NEW is based on data transformation and parametric process monitoring (cf., its description in Section 2). From this example, it can be seen that the proposed chart NEW would not lose much reliability in various cases when the normality and "data independence" assumptions are violated by using data decorrelation, data transformation and parametric process monitoring, in comparison with its peers, while the benefit to use NEW for effective detection of process distributional shifts is quite profound, as will be seen from the examples below.

Besides the actual ARL_0 values, we also use the false alarm rate (FAR) to compare the IC performance of different methods, where FAR is defined to be the probability that the process under monitoring is declared to be out-of-control (OC) when it is actually IC. Since all control charts will eventually give a false signal in any IC simulation run, here FAR of a given chart is defined to be the proportion of IC simulation runs in which the chart gives a false signal within the first 50 observation times. Namely, it is defined to be $P(RL \leq 50)$ when the process is IC. When $m_0 = 300$ and all other setups are the same as those in the example of Table 1, the FAR results are presented in Table 2. From the table, it can be seen that the charts SS-MEWMA, SR-MEWMA and SR-CPD have reasonable performance only in cases when their model assumptions are valid (e.g., Cases I and II for the nonparametric charts SR-MEWMA and SR-CPD). As a comparison, the charts XQ-CUSUM and NEW both perform well in all cases considered, and NEW is slightly better than XQ-CUSUM in this example.

Table 1: Estimated ARL_0 values and their standard errors (in parentheses) of different control charts when their nominal ARL_0 values are fixed at 200, and the IC sample size m_0 changes among $\{100, 200, 300, 500\}$.

Cases	Methods	$m_0 = 100$	200	300	500
I	SS-MEWMA	188(4.36)	195(3.92)	200(3.09)	199(2.50)
	SR-MEWMA	188(3.68)	188(3.22)	194(2.55)	193(1.87)
	SR-CPD	206(3.89)	206(3.32)	212(2.81)	201(2.03)
	XQ-CUSUM	206(4.27)	213(3.85)	209(3.95)	202(3.09)
	NEW	169(6.91)	174(6.39)	190(6.64)	202(5.45)
II	SS-MEWMA	162(4.88)	160(3.82)	175(4.56)	182(4.31)
	SR-MEWMA	187(3.32)	189(3.15)	188(2.28)	190(1.77)
	SR-CPD	210(3.63)	202(3.31)	210(2.71)	201(1.91)
	XQ-CUSUM	178(4.36)	181(3.86)	209(4.41)	201(3.30)
	NEW	178(6.32)	176(6.82)	198(6.01)	201(5.35)
III	SS-MEWMA	68.8(1.76)	67.4(1.27)	73.2(1.37)	74.6(1.11)
	SR-MEWMA	64.7(1.02)	62.9(0.91)	65.7(0.61)	65.9(0.50)
	SR-CPD	65.0(1.01)	62.8(0.82)	65.7(0.61)	65.4(0.48)
	XQ-CUSUM	170(4.25)	178(3.99)	205(4.31)	208(3.32)
	NEW	169(6.96)	171(6.74)	193(6.86)	198(5.24)
IV	SS-MEWMA	68.7(1.76)	67.3(1.27)	73.2(1.36)	74.6(1.10)
	SR-MEWMA	64.7(1.07)	62.8(0.91)	65.7(0.60)	65.9(0.49)
	SR-CPD	65.2(1.01)	63.0(0.84)	65.5(0.64)	65.3(0.48)
	XQ-CUSUM	161(4.24)	177(3.92)	205(4.21)	207(3.30)
	NEW	168(6.71)	171(6.73)	194(6.85)	196(5.23)

Table 2: Estimated FAR values of different control charts when $m_0 = 300$ and other setups are the same as those in the example of Table 1.

Cases	SS-MEWMA	SR-MEWMA	SR-CPD	XQ-CUSUM	NEW
I	0.173	0.179	0.165	0.201	0.198
II	0.215	0.191	0.169	0.196	0.191
III	0.474	0.513	0.505	0.222	0.196
IV	0.473	0.514	0.505	0.223	0.195

Evaluation of the OC performance: Next, we evaluate the OC performance of the related charts in case when $m_0 = 300$. In order to make the comparison more meaningful, we intentionally

adjust the control limits of different control charts based on simulations so that their actual ARL_0 values equal the nominal ARL_0 value of 200 in all cases considered. In the next simulation example, it is assumed that all quality variables have a same shift at the beginning of online process monitoring with the shift size δ changing from -1.5 to 1.5 with a step of 0.25. Because different control charts have different procedure parameters (e.g., the weighting parameters of SS-MEWMA, SR-MEWMA and NEW) and their performance may not be comparable if their parameters are set to be the same, here we compare their optimal OC performance to make the comparison fair. Namely, to detect a given shift by a chart, the related procedure parameter is chosen by minimizing the ARL_1 value of the chart while maintaining its ARL_0 value at 200. The resulting ARL_1 value is called optimal ARL_1 value hereafter.

The results of the computed optimal ARL_1 values of the five charts are presented in Figure 1 in Cases I-IV when $m_0 = 300$. From the figure, we can have the following conclusions. i) All charts perform reasonably well in Case I since their model assumptions are all satisfied. ii) The chart SS-MEWMA performs the best in Case I when its normality and "data independence" assumptions are valid, but is less effective in Cases II-V when one or both of these assumptions are invalid. iii) The charts XQ-CUSUM and NEW perform better than the other three charts in Cases III and IV when IC process observations are serially correlated. iv) Between the two charts XQ-CUSUM and NEW, NEW has a better performance when the shift size is relatively large (e.g., $\delta \geq 0.75$), and is slightly worse when the shift size is small.

The proposed chart NEW depends on the initial estimates $\{\widehat{F}_j^{(0)}(x), 1 \leq j \leq p\}$ of the cdf's of the quality characteristics under monitoring, whose variability would be large when the IC sample size m_0 is small. To study the impact of m_0 on the OC performance of NEW and other related charts, next we consider an example with $m_0 = 1000$ and all other setups being the same as those in the example of Figure 1. The results of the computed optimal ARL_1 values of the five charts are presented in Figure 2. From the figure, it can be seen that similar conclusions to those in the previous example can be made here, although the advantage to use the proposed method NEW seems more obvious in this example, since its performance is the best or close to the best among all five methods in different cases considered, even when the shift size is small. As a summary, the simulation examples presented in this part confirm that the proposed chart NEW is more effective for online process monitoring in most cases considered, in comparison with its peers.

To study the impact of the IC sample size m_0 on the OC performance of the proposed method

NEW, next we consider an example in which m_0 can take the values of 100, 200, 300, 500 and 1,000, and other setups remain the same as those in the example of Figure 1. Its calculated optimal ARL_1 values are presented in Figure 3. From the plots in the figure, it can be seen that i) the OC performance of NEW is better when m_0 is larger, and ii) its OC performance is reasonably stable when $m_0 \geq 300$ in most cases considered in this example.

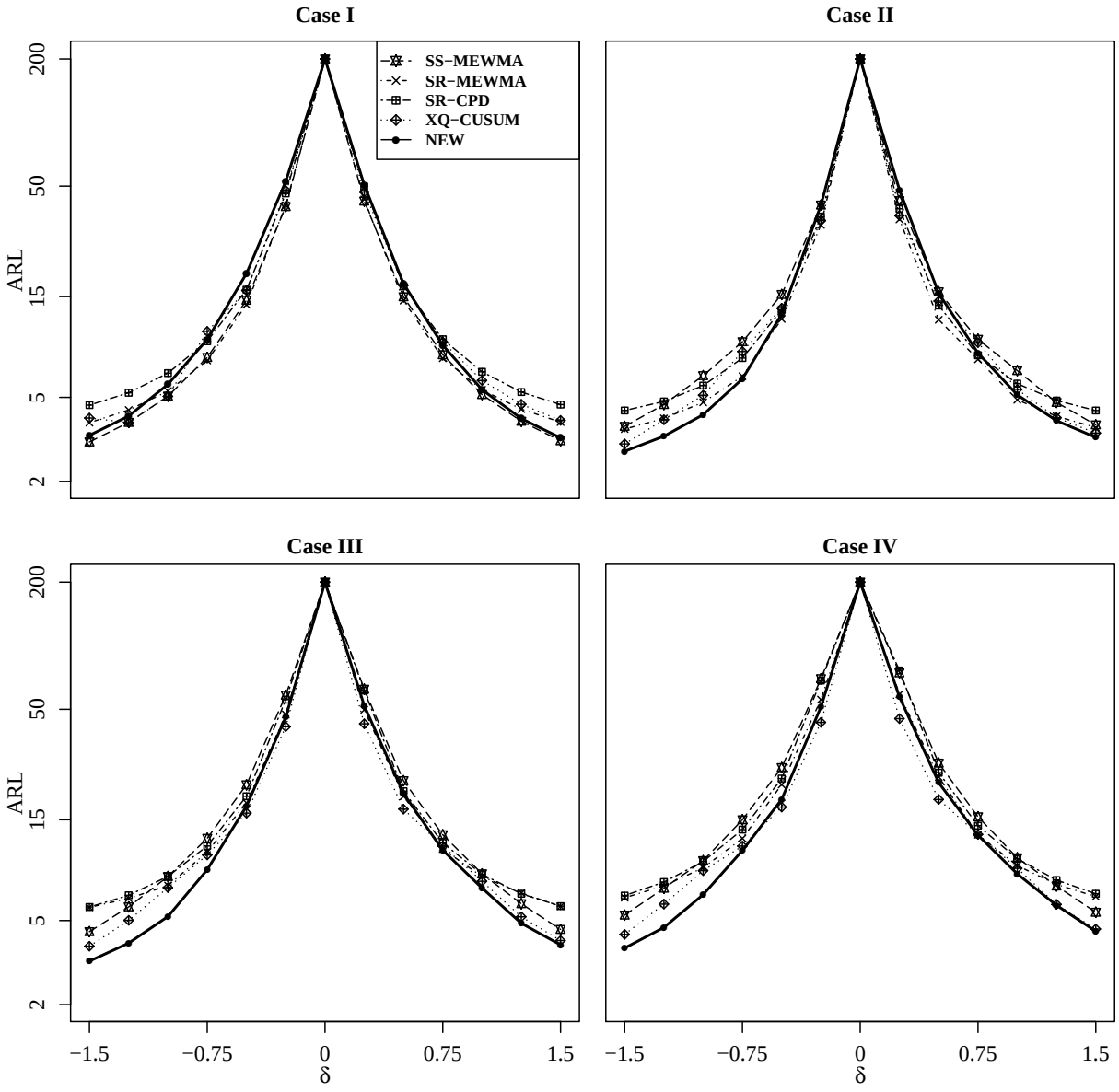

Figure 1: Optimal ARL_1 values of the five control charts when their nominal ARL_0 values are fixed at 200, $p = 3$, $m_0 = 300$, and all quality characteristics have the same shift with the shift size δ changing among ± 0.25, ± 0.5, ± 0.75, ± 1, ± 1.25 and ± 1.5.

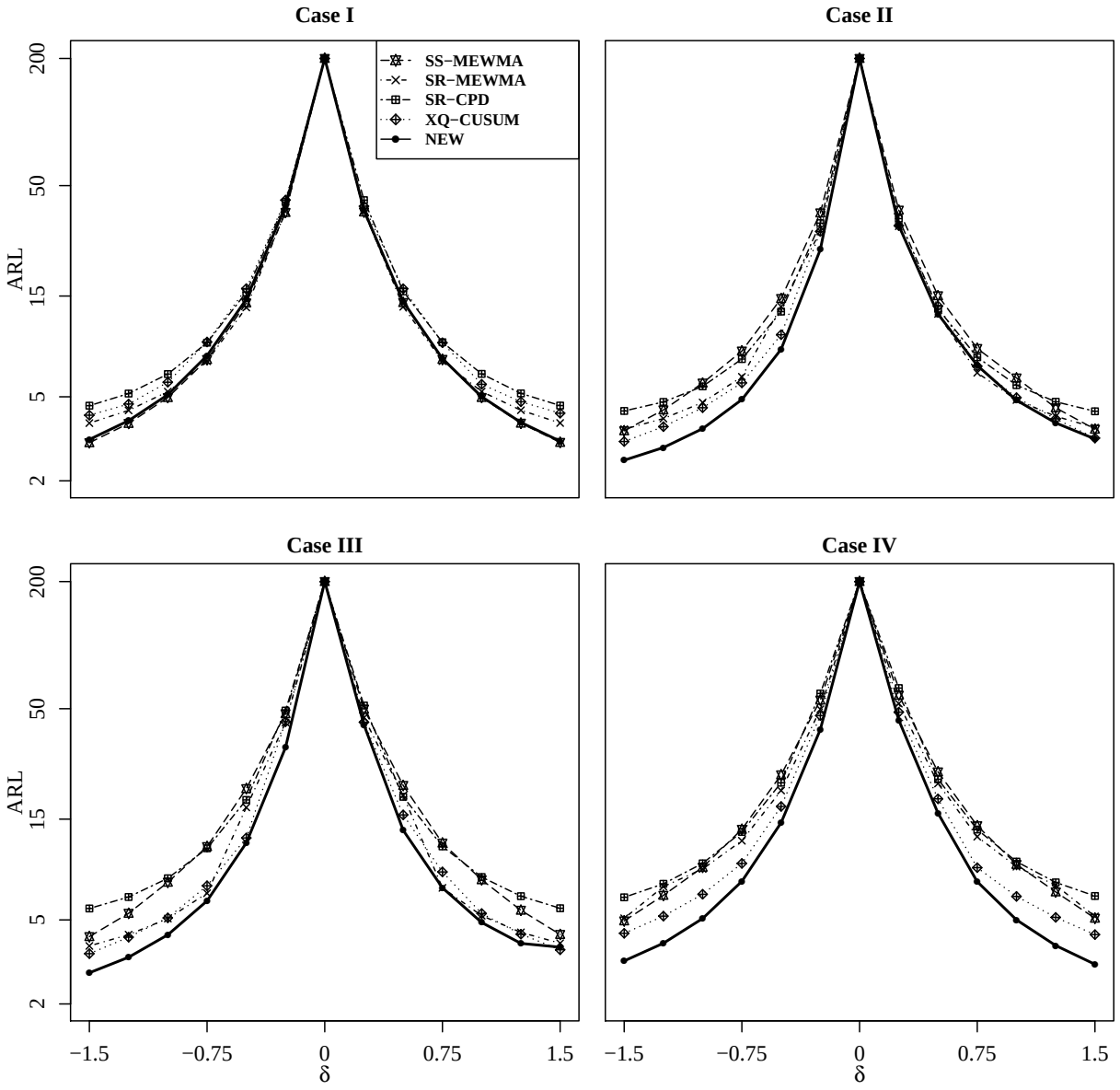

Figure 2: Optimal ARL_1 values of the five control charts when their nominal ARL_0 values are fixed at 200, $p = 3$, $m_0 = 1000$, and all quality characteristics have the same shift with the shift size δ changing among ± 0.25, ± 0.5, ± 0.75, ± 1, ± 1.25 and ± 1.5.

4 A Case Study

In this section, a real-life dataset from a mining process is used to demonstrate the application of the proposed chart NEW, which can be downloaded from the web page of Kaggle with the link https://www.kaggle.com/edumagalhaes/quality-prediction-in-a-mining-process. The flotation

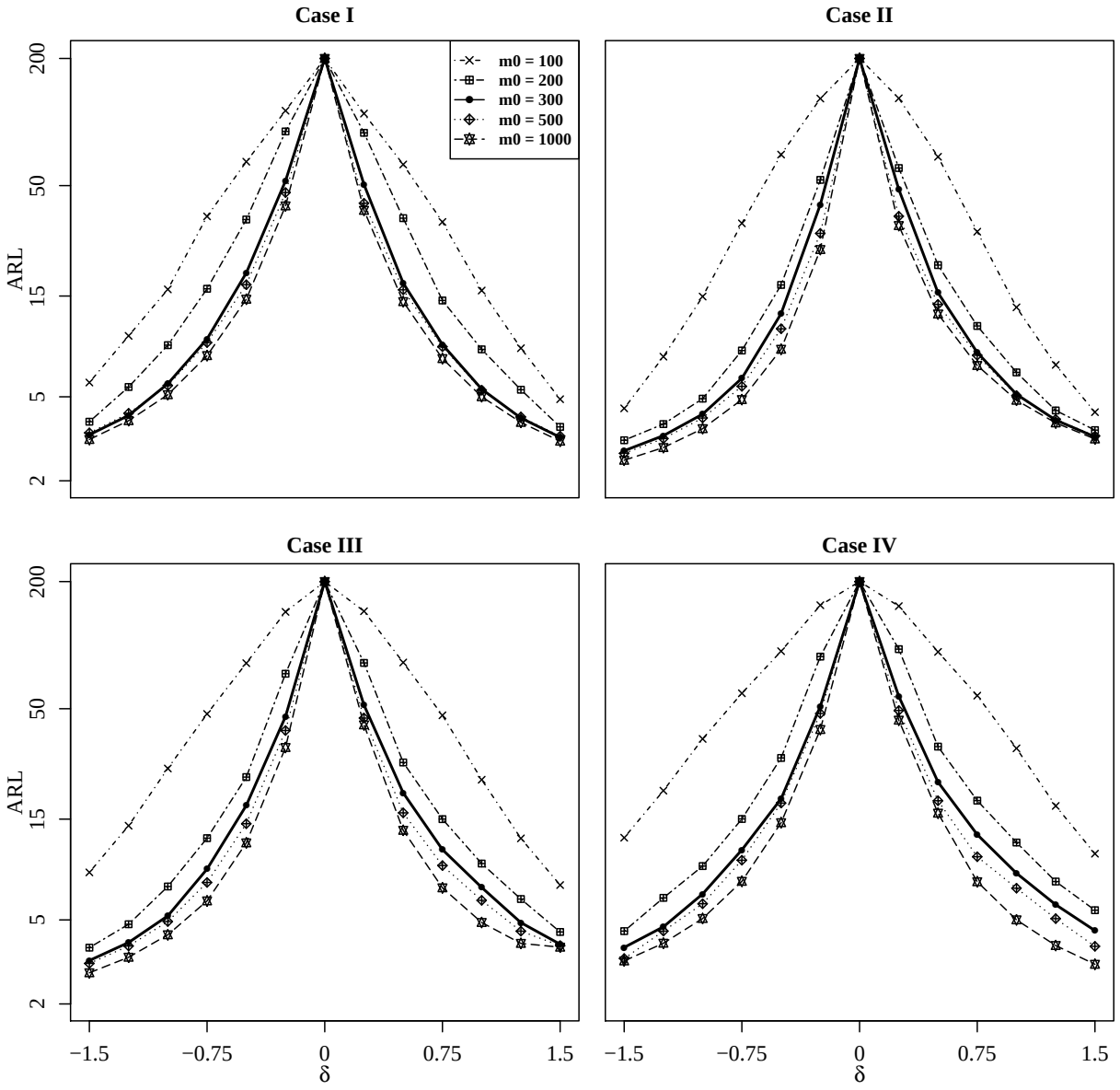

Figure 3: Optimal ARL_1 values of the chart NEW when its nominal ARL_0 value is fixed at 200, m_0 changes among 100, 200, 300, 500, and 1000, and other setups are the same as those in Figure 1.

method is often used in mineral processing to concentrate ores by separating hydrophobic materials from hydrophilic materials (cf., Crawford and Quinn 2017). See Figure 4 for a demonstration. Online monitoring of data streams collected from the flotation process is especially important because they would affect the impurity of ore concentrate if something unusual happens.

The data used here contain observations of three major characteristics of a flotation process: ore

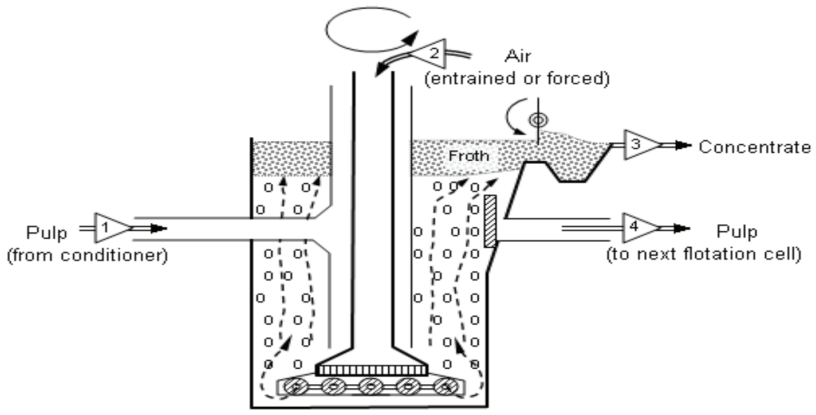

Figure 4: Demonstration of a flotation process in mineral processing to concentrate ores.

pulp flow, flotation air flow, and flotation level. The original data of these three variables are shown in Figure 5. From the figure, it seems that the first 300 observations are quite stable, and thus they are used as the IC data. The remaining 100 observations are used for online process monitoring. For the IC data, we first check for serial data correlation using the Ljung–Box test. The p-values of this test for the three quality characteristics are $< 2.2 \times 10^{-16}$, $< 2.2 \times 10^{-16}$, and 9.99×10^{-16}, respectively. Thus, there is a significant autocorrelation in the observed data of all three variables. The Augmented Dickey-Fuller (ADF) test for stationarity of the autocorrelation gives p-values of < 0.01 for all three variables, which implies that the stationarity assumption is valid in this case. To check the normality assumption for the data, the Shapiro test is performed, and its p-values for the three quality characteristics are 0.006, 8.77×10^{-5}, and 8.93×10^{-1}, respectively. Thus, the normality assumption is significantly violated for all three variables. Therefore, the IC data have a significant stationary serial data correlation, and a non-normal distribution in this example.

Next, we apply the five control charts SS-MEWMA, SR-MEWMA, SR-CPD, XQ-CUSUM and NEW to this data for online process monitoring starting from the 301^{st} observation time. In all control charts, the nominal ARL_0 values is fixed at 200, and their control limits are computed in the same way as that in the simulation study for evaluating their IC performance in Section 3. The five control charts are shown in Figure 6. From the plots in the figure, the charts SS-MEWMA, SR-MEWMA and SR-CPD give signals at many observation times. Since their model assumptions of "data independence" and normality (for SS-MEWMA) are violated in this example, their results may not be reliable, as discussed in Section 3. The charts NEW and XQ-CUSUM give their first

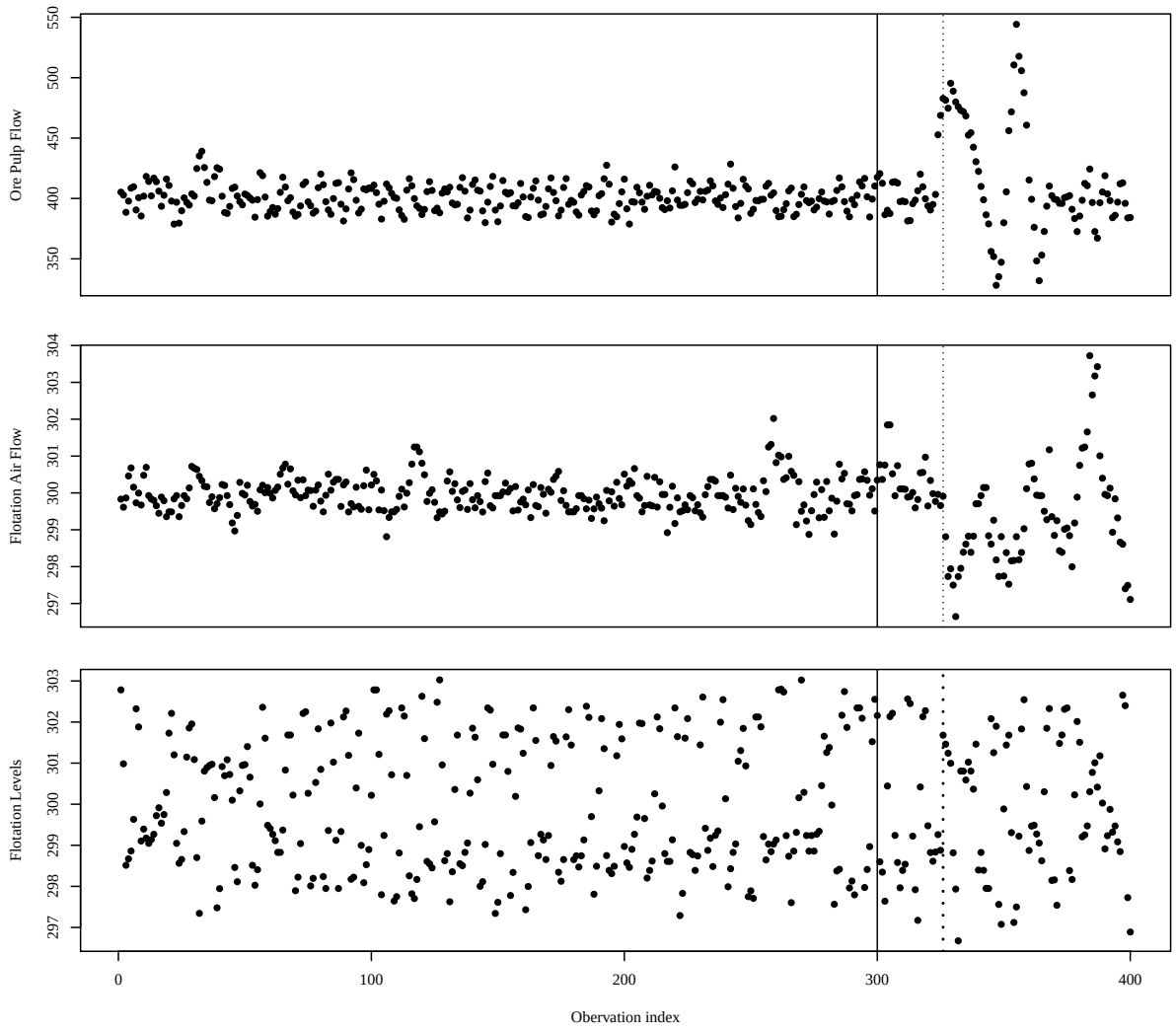

Figure 5: Original observations of three variables in the mining process. The solid vertical line in each plot separates the initial IC data from the data for online process monitoring, and the dashed line indicates the signal time of our method NEW.

signals at the 326^{th} and 341^{th} observation times, respectively. By checking the original process observations shown in Figure 5, the signal of NEW indicates the start of a systematic process mean shift well.

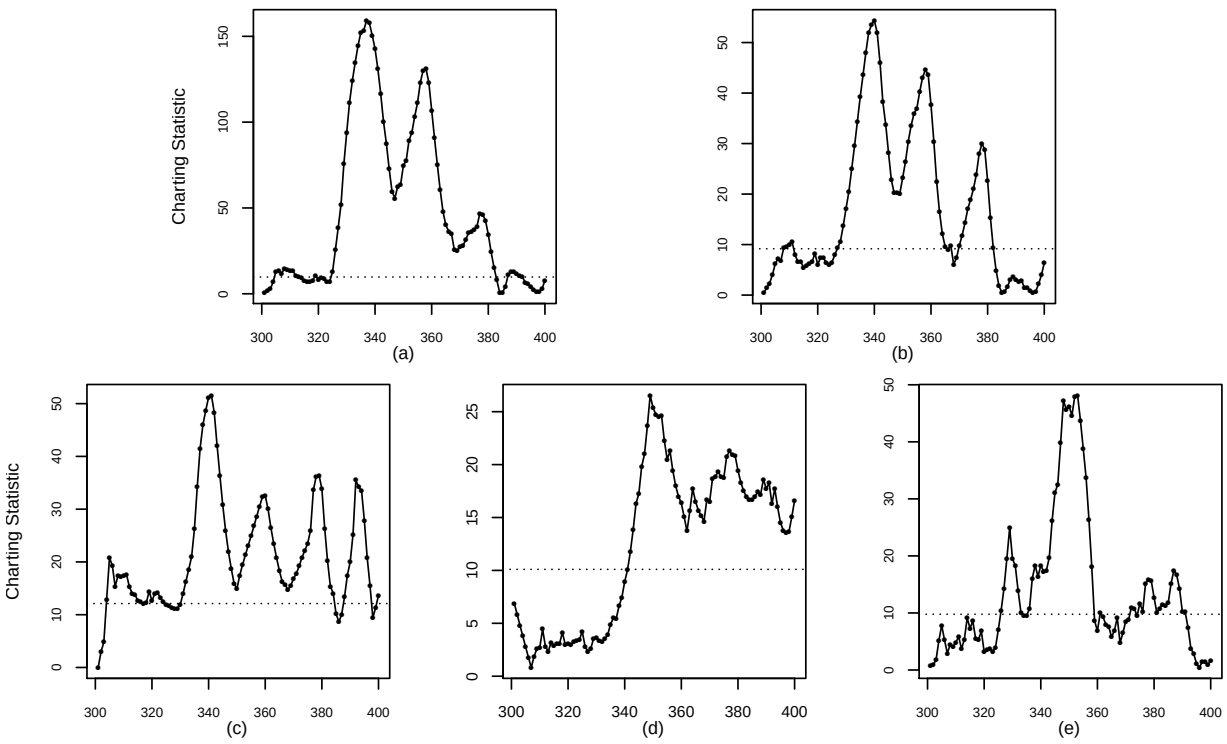

Figure 6: Control charts SS-MEWMA (plot (a)), SR-MEWMA (plot (b)), SR-CPD (plot (c)), XQ-CUSUM (plot (d)) and NEW (plot (e)) for monitoring the last 100 process observations shown in Figure 4. In each plot, the horizontal dotted line denotes the control limit of the related control chart.

5 Concluding Remarks

Recently many nonparametric multivariate SPC charts have been developed for handling cases when the IC process distribution do not have a parametric form. Most existing nonparametric multivariate SPC charts are based on data ordering and/or data categorization. Thus, a substantial amount of information in the original process observations would be lost, which would negatively affect the effectiveness of these charts for online process monitoring. To overcome this limitation, we have proposed an alternative approach to handle the nonparametric multivariate SPC problem in this paper. Instead of data ordering and/or data categorization, our proposed method is based on data decorrelation, estimation of the IC process distribution, and data transformation. Numerical studies presented in Sections 3 and 4 show that it performs well in various cases considered. However, there are still some issues about the proposed method that need to be addressed in the future research. For instance, when the number of quality variables is large, the required initial IC

sample size also should also be large in order to have a reliable performance of the related chart. In such cases, some variable selection approaches might be helpful (cf. Zou and Qiu 2009). Also, the control limit of the proposed method is determined by the Monte Carlo simulation based on the assumption that the transformed observations $\{\mathbf{Z}_n, n \geq 1\}$ are i.i.d. with a normal IC distribution. Although the simulation results in Section 3 have shown that the proposed chart using the control limit determined in that way performs reasonably well in all cases considered there, some theoretical justifications are needed. All these issues will be studied carefully in our future research.

Acknowledgments: The authors thank the editor and a referee for some constructive comments and suggestions which improved the quality of the paper greatly. This research is supported in part by the NSF grant DMS-1914639.

References

Apley, D. W., and Lee, H. C. (2008), "Robustness comparison of exponentially weighted moving-average charts on autocorrelated data and on residuals," *Journal of Quality Technology*, **40**, 428–447.

Capizzi, G., and Masarotto, G. (2008), "Practical design of generalized likelihood ratio control charts for autocorrelated data," *Technometrics*, **50**, 357–370.

Capizzi, G., and Masarotto, G. (2016), "Efficient control chart calibration by simulated stochastic approximation," *IIE Transactions*, **48**, 57–65.

Chakraborti, S., and Graham, M.A. (2019), "Nonparametric (distribution-free) control charts: An updated overview and some results," *Quality Engineering*, **31**, 523–544.

Crawford, C.B., and Quinn, B. (2017), "Microplastic separation techniques," *Microplastic Pollutants*, 203–218.

Hawkins, D.M. (1987), "Self-starting cusums for location and scale," *The Statistician*, **36**, 299–315.

Hawkins, D.M., and Maboudou-Tchao, E.M. (2007), "Self-starting multivariate exponentially weighted moving average control charting," *Technometrics*, **49**, 199–209.

Hawkins, D.M., and Olwell, D.H. (1998), *Cumulative Sum Charts and Charting for Quality Improvement*, New York: Springer-Verlag.

Hawkins, D.M., Qiu, P., and Kang, C.W. (2003), "The changepoint model for statistical process control," *Journal of Quality Technology*, **35**, 355–366.

Higham, N.J. (1988), "Computing a nearest symmetric positive semidefinite matrix," *Linear Algebra and its Applications*, **103**, 103–118.

Holland, M.D., and Hawkins, D.M. (2014), "A control chart based on a nonparametric multivariate change-point model," *Journal of Quality Technology*, **46**, 63–77.

Lee, H. C., and Apley, D. W. (2011), "Improved design of robust exponentially weighted moving average control charts for autocorrelated processes," *Quality and Reliability Engineering International*, **27**, 337–352.

Li, J. (2021), "Nonparametric adaptive CUSUM chart for detecting arbitrary distributional changes," *Journal of Quality Technology*, **53**, 154–172.

Li, J., Zhang, X., and Jeske, D.R. (2013), "Nonparametric multivariate CUSUM control charts for location and scale changes," *Journal of Nonparametric Statistics*, **25**, 1–20.

Lowry, C.A., Woodall, W.H., Champ, C.W., and Rigdon, S.E. (1992), "A multivariate exponentially weighted moving average control chart," *Technometrics*, **34**, 46–53.

Montgomery, D.C. (2012), *Introduction to Statistical Quality Control,* New York: John Wiley & Sons.

Page, E.S. (1954), "Continuous inspection scheme," *Biometrika*, **41**, 100–115.

Qiu, P. (2008), "Distribution-free multivariate process control based on log-linear modeling," *IIE Transactions*, **40**, 664–677.

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

Qiu, P. (2018), "Some perspectives on nonparametric statistical process control," *Journal of Quality Technology*, **50**, 49–65.

Qiu, P., and Hawkins, D.M. (2001), "A rank based multivariate CUSUM procedure," *Technometrics*, **43**, 120–132.

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

Roberts, S.V. (1959), "Control chart tests based on geometric moving averages," *Technometrics*, **1**, 239–250.

Shewhart, W.A. (1931), *Economic Control of Quality of Manufactured Product,* New York: D. Van Nostrand Company.

Xue, L., and Qiu, P. (2021), "A nonparametric CUSUM chart for monitoring multivariate serially correlated processes," *Journal of Quality Technology*, **53**, 396–409.

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

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